UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

WASHINGTON, DC 20460

September 28, 2001

OFFICE OF
ENVIRONMENTAL INFORMATION

MEMORANDUM

SUBJECT: Review of *Guidance on Data Quality Indicators (EPA QA/G-5i)*

FROM: Nancy Wentworth, Director */s/ Nancy Wentworth*
Quality Staff (2811R)

TO: Peer Review Panel

Attached is the Quality Staff Peer Review draft of the document *Guidance on Data Quality Indicators (EPA QA/G-5i)*. Your help in supplying review comments would be greatly appreciated. This document is intended to assist a wide audience of environmental analysts, managers, and decision makers in the collection and assessment of environmental data. Although the document does not emphasize statistical theory and formulae, basic equations to calculate various indicators are included.

The document contains six chapters: an introduction, an overview of Data Quality Indicators, Data Quality Indicators generally related to measurements, Data Quality Indicators generally related to sampling, the extension of indicators beyond precision, accuracy, representativeness, comparability, completeness, and sensitivity, and finally a brief discussion on how these are integrated into a project's life-cycle.

As a reviewer of this document, please give attention to the organizational aspects of this guidance as well as the indicator discussion aspects of the document. Some specifics questions the reviewer could address include:

Overall

- Is this guidance written in plain English to ensure that the information is accessible to those who can most effectively utilize it to improve data collection efforts?
- Should there be a preamble similar to the one in *Guidance on the Data Quality Objectives Process (EPA QA/G-4)* or can this be discarded as the document is of a specialized nature intended for a technical audience?

- Are there sufficient statistical formulae included to allow for adequate use of the guidance? If not, what would you recommend?

- Is too much time devoted to explanations of the indicators?

Specific Chapters

- Chapter 2 -- Should a discussion on geostatistics be included in this chapter or is the discussion on sampling units (Section 2.3) sufficient?

- Chapter 3 -- In Section 3.1 reference to the use of quality control data to establish Data Quality Indicators is made but not investigated in length. Do you think this is sufficient to make the point or should the section be expanded by further examples and calculations?

- Chapter 3 -- Figure 3-5 uses the concept of triangles to demonstrate the components of variance. Does this work?

- Chapter 3 -- Should the statistical understanding of "comparability" be investigated even though this is not really a statistical document?

- Chapter 3 -- Sensitivity is discussed almost solely in terms of detection limits. Should instrument resolution be included and do you have some suggestions for references?

- Chapter 3, Section 3.5 -- The section has information on different types of detection limits. Is this sufficient bearing in mind that a complete book on the topic could be written on the subject? Should sensitivity contain information on how to calculate different types of detection limits through actual examples or is the discussion of various methods to determine detection limit sufficient?

- Chapter 4, Section 4.1 -- Should a more in-depth discussion of Gy's Theory of Sampling be included in this section?

- Chapter 5 -- Are there other indicators that should be included?

Your comments and suggestions for improvement, together with references to a particular section of a publication or handbook for additional or relevant topics would be appreciated.

The attachment is a Adobe-Acrobat version, if you would prefer a paper version, please let me know and a copy will be sent to you. We would appreciate hand-written comments on the original document rather than a formal line-by-line separate response to this request. All comments are requested from minor typographical elements to major changes in direction. Any comments should be emailed to John Warren at warren.john@epa.gov or faxed to 202-565-2441 by November 30, 2001.

Attachment

Guidance on Data Quality Indicators

EPA QA/G-5i

Quality Staff
Office of Environmental Information
United States Environmental Protection Agency
Washington, D.C. 20460

PEER REVIEW DRAFT

September 2001

FOREWORD

The U.S. Environmental Protection Agency (EPA) has developed an Agency-wide program of quality assurance for environmental data. An understanding of the principal data quality indicators (DQIs) and where they fit in the project life cycle, is essential for designing and implementing a quality system. This guidance document, *Guidance on Data Quality Indicators*, provides an overview of the individual DQIs, and how they fit into the overall project life cycle.

This document is one of the *U.S. Environmental Protection Agency Quality System Series* documents. These documents describe the EPA policies and procedures for planning, implementing, and assessing the effectiveness of the Quality System. As required by EPA Manual 5360 A1 (May 2000b), this document is valid for a period of up to five years from the official date of publication. After five years, this document will be reissued without change, revised, or withdrawn from the *U.S. Environmental Protection Agency Quality System Series* documents. Questions regarding this document or other *Quality System Series* documents should be directed to the Quality Staff at:

U.S. EPA
Quality Staff (2811R)
1200 Pennsylvania Ave., NW
Washington, DC 20460
Phone: (202) 564-6830
Fax: (202) 565-2441
E-mail: quality@epa.gov

Copies of the *Quality System Series* documents may be obtained from the Quality Staff directly or by downloading them from its Home Page:

www.epa.gov/quality

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EXAMPLES

LIST OF ACRONYMS

AML	alternative minimum level
ANSI	American National Standards Institute
CFR	Code of Federal Regulations
CV	coefficient of variation
DQI	data quality indicator
DQO	data quality objective
ELCD	electrolytic conductivity detector
GC	gas chromatograph
ICP	inductively coupled plasma
IDE	interlaboratory detection estimate
IDL	instrument detection limit
IQCL	instrument quality control level
IQE	interlaboratory quantitation estimate
IUPAC	International Union for Pure and Applied Chemistry
LOD	limit of detection
LOQ	limit of quantification
LRL	laboratory reporting level
LT-MDL	long-term method detection level
MDL	method detection limit
MQCL	method quality control level
MQO	measurement quality objective
MS	mass spectrometer
PARCCS	precision, accuracy (bias), representativeness, comparability, completeness, and sensitivity
ppb	parts per billion
ppm	parts per million
PQL	practical quantitation limit
QA	quality assurance
QC	quality control
RCRA	Resource Conservation and Recovery Act
RDL	reliable detection level
RL	reporting limit
RPD	relative percent difference
RSD	relative standard deviation
SIM	selected ion monitoring
SVOC	semivolatile organic compound
TOC	total organic compounds
VOC	volatile organic compound
XRF	x-ray fluorescence

CHAPTER 1

INTRODUCTION

1.1 PURPOSE OF DOCUMENT

This document provides guidance to the environmental science community on various facets of data quality indicators (DQIs), and their role throughout the project life cycle. DQIs are quantitative and qualitative measures of principal quality attributes, including precision, accuracy, representativeness, comparability, completeness, and sensitivity (PARCCS). Historically, DQIs sometimes have been incorrectly equated with data quality objectives (DQOs), which are specifications for decision making. This guidance document will clear up this confusion, and help users better understand what DQIs are, why they are important, and how they can be used.

While this document does not emphasize the use of statistical theory and formulae, basic equations used to calculate various indicators are provided, along with examples, so that this document can serve as a stand-alone guide to DQIs.

1.2 THE EPA QUALITY SYSTEM

A quality system is a structured and documented management system describing the policies, objectives, principles, organizational authority, responsibilities, accountability, and implementation plan of an organization for ensuring quality in its work processes, products, and services. A quality system provides the framework for planning, implementing, documenting, and assessing work performed by the organization for carrying out required quality assurance (QA) and quality control (QC) activities.

Since 1979, U.S. Environmental Protection Agency (EPA) policy has required participation in an Agency-wide quality system by all EPA organizations (i.e., office, regions, national centers, and laboratories) supporting intramural environmental programs, and by non-EPA organizations performing work funded by EPA through extramural agreements. The EPA Quality System operates under the authority of Order 5360.1 A2, *Policy and Program Requirements for the Mandatory Agency-wide Quality System* (U.S. EPA, 2000a). The implementation requirements for the Order for EPA organizations are provided in Order 5360 A1, *EPA Quality Manual for Environmental Programs* (U.S. EPA, 2000b).

1.3 SCOPE AND BACKGROUND

Figure 1-1 depicts the project life cycle of environmental data in EPA's Quality System. This guidance document is intended to provide users with a better understanding of data quality indicators, as well as the specifications regarding these indicators that are required inputs to the QA Project Plan. It also provides useful information to augment the concepts and procedures

discussed in *Guidance on Sampling Design* (EPA QA/G-5s) (U.S. EPA, 2000e) because many of the indicators are important inputs to the design process, and *Guidance on Data Quality Assessment* (EPA QA/G-9) (U.S. EPA, 2000c) because many of the indicators provide useful information during the assessment of data adequacy.

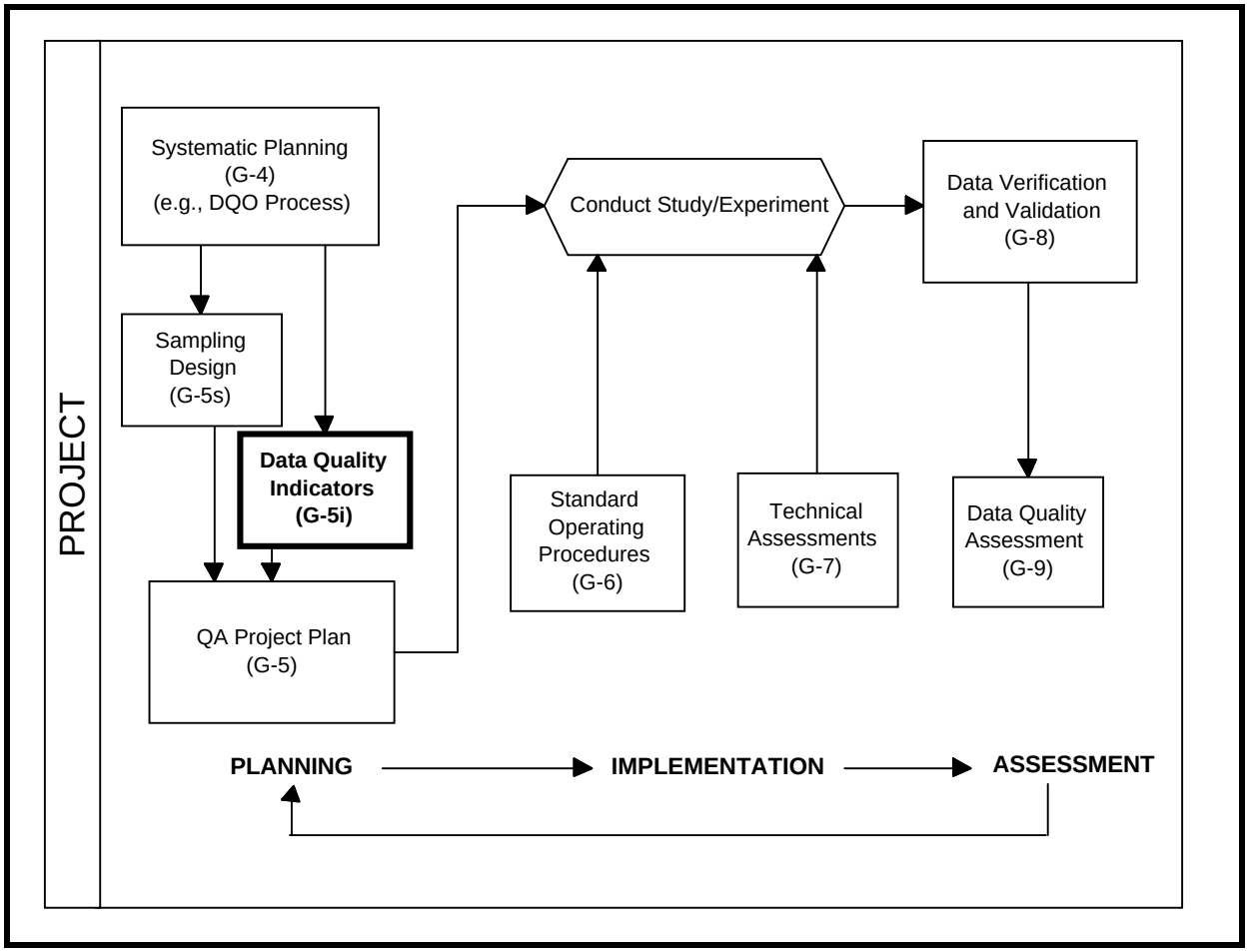


Figure 1-1. Life Cycle of Data in the EPA Quality System

1.4 INTENDED AUDIENCE

This document is aimed at a wide audience of individuals involved in the collection and assessment of environmental data. It is designed to support the efforts of:

- those responsible for preparing or reviewing project planning documents such as QA Project Plans,
- those involved in implementing data collection activities in the field or laboratory,

- data validators and data quality assessment (DQA) analysts, and
- those who use environmental data in decision-making.

This guidance may be used by any organization implementing environmental data collection programs encompassing field sampling and analytical activities.

1.5 SPECIFIC DEFINITIONS

For purposes of consistency, the following terms are used through this document. These definitions do not constitute the Agency's official use of terms for regulatory purposes and should not be construed to alter or supplant other terms in use.

precision - the measure of agreement among repeated measurements of the same property under identical, or substantially similar, conditions.

bias - systematic or persistent distortion of a measurement process that causes errors in one direction.

accuracy - a measure of the overall agreement of a measurement to a known value.

representativeness - the measure of the degree to which data accurately and precisely represent a characteristic of a population, parameter variations at a sampling point, a process condition, or an environmental condition.

comparability - the qualitative term that expresses the measure of confidence that two or more data sets can contribute to a common analysis.

completeness - a measure of the amount of valid data obtained from a measurement system.

sensitivity - the capability of a method or instrument to discriminate between measurement responses representing different levels of the variable of interest.

The discussions in the subsequent chapters will amplify on these brief definitions.

1.6 PERIOD OF APPLICABILITY

Based on the *Quality Manual* (U.S. EPA, 2000b), this document will be valid for a period of five years from the official date of publication. After five years, this document will either be reissued without modification, revised, or removed from the EPA Quality System.

1.7 ORGANIZATION OF THIS DOCUMENT

An overview, including general definitions for each DQI, is provided in Chapter 2. This chapter provides a framework for discussing the various components of total study error, and introduces some central concepts that are important for understanding the discussions that follow. For example, Chapter 2 differentiates between measurement and sampling error, and recognizes that the problem of measuring characteristics of interest within a sampling unit is complicated by small-scale variability and the process of acquiring one or more specimens from that unit. Chapter 3 presents DQIs generally related to environmental measurements (precision, accuracy, comparability, and sensitivity), while Chapter 4 presents DQIs generally related to sampling (representativeness and completeness). Chapter 5 includes an overview of additional DQIs beyond the PARCCS parameters. Chapter 6 provides a brief discussion of how the DQIs are integrated and considered during the planning, implementation, and assessment portions of the project life cycle. Examples are interspersed throughout to illustrate important concepts and demonstrate the calculation of the quantitative indicators.

CHAPTER 2

OVERVIEW OF DATA QUALITY INDICATORS

2.1 GENERAL DEFINITIONS FOR QUALITY ATTRIBUTES MEASURED BY DQIs

Data quality is a very general term. In its broadest sense, data quality is a measure of the degree of acceptability or utility of data for a particular purpose. To simplify the way data quality is examined, and to facilitate communication about data quality issues, certain data quality attributes can be defined and measured. The principal quality attributes important to environmental studies are precision, bias, representativeness, comparability, completeness, and sensitivity. These six quality attributes are also referred to by the acronym PARCCS, with the "A" in PARCCS referring to accuracy instead of bias. [Note: This substitution of "A" for "B" occurs because some analysts believe accuracy and bias are synonymous, and PARCCS is a historically recognized and familiar acronym. Accuracy is actually comprised of random error (precision) and systematic error (bias), and these indicators are discussed separately.] DQIs are qualitative and quantitative measures of data quality attributes; they are not themselves data quality attributes. For example, precision is not a DQI, but a DQI can certainly be defined to provide some measure of precision. The DQI is just as the name suggests: an indicator of an underlying data quality attribute.

DQIs for precision, bias, and sensitivity can be defined and measured in quantitative terms; representativeness, comparability, and completeness have more qualitative definitions. Establishing acceptance criteria for the DQIs sets quantitative goals for the quality of data generated in the analytical measurement process. Individual measurement quality objectives (MQOs), are "acceptance criteria" for the quality attributes measured by project DQIs. The relationship among these quality terms (DQOs, attributes, DQIs, and MQOs) is presented in Table 2-1. The MQOs reflect the outcome of the DQO planning process (see Figure 2-1), but are not synonymous with project DQOs. The relationship between MQOs and DQOs is discussed throughout this document in more detail, with a summary provided in Chapter 6.

Table 2-1. Relationship Among Quality Terms

DQOs	Qualitative and quantitative study objectives for collection of data
Attributes	Qualitative and quantitative characteristics of the collected data
DQIs	Indicators of the quality attributes
MQOs	Acceptance thresholds or goals for the data, usually based on individual DQIs

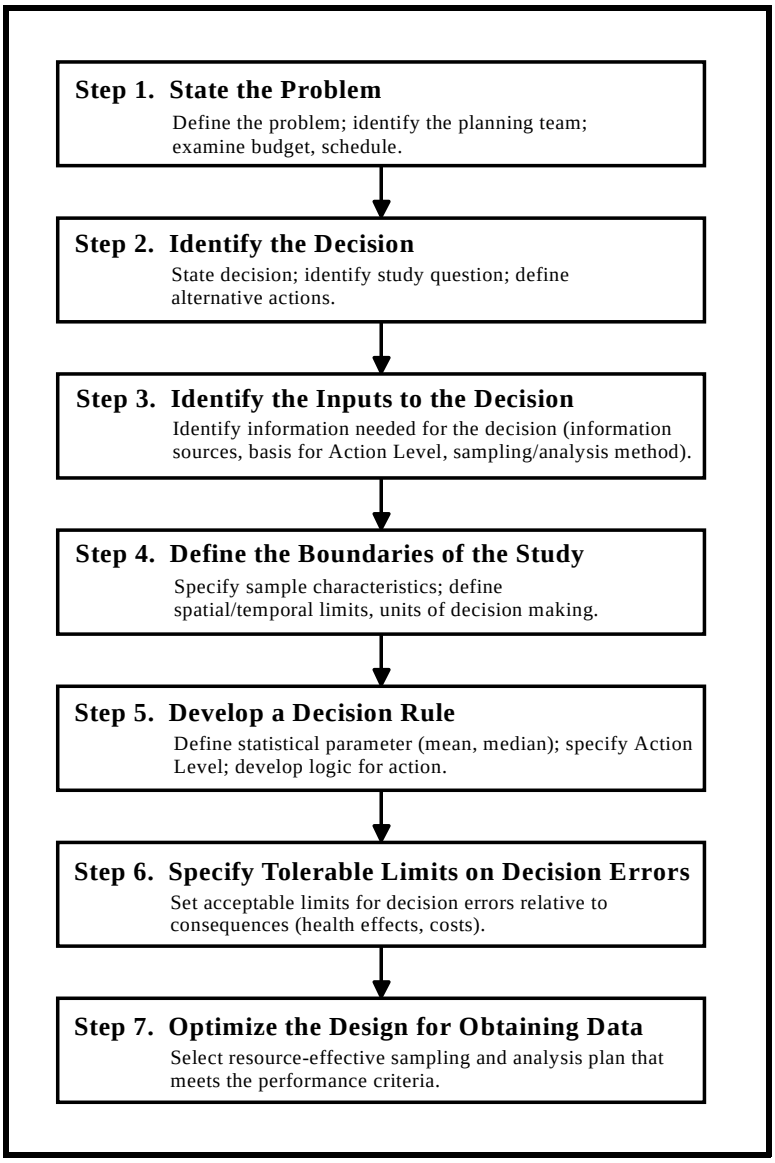


Figure 2-1. The Data Quality Objectives Process

DQIs customarily are applied only to the laboratory measurement processes, but, as discussed later, DQIs can be chosen to capture the effects of other important study processes and procedures on the overall quality of study data. The next few paragraphs briefly review the definitions for data quality attributes commonly monitored with DQIs.

Precision is the measure of agreement among repeated measurements of the same property under identical, or substantially similar, conditions. For environmental studies, the property is typically a concentration of a contaminant, but any physical measurement has error, and therefore the precision of that measurement can be examined. Random errors or fluctuations in the measurement process always result in some range of values of these repeated

measurements. A precision DQI is a quantitative indicator of the dispersion generated from these random errors.

Bias is systematic or persistent distortion of a measurement process that causes errors in one direction. Bias may originate from sources such as calibration errors, response factor shifts, unaccounted-for interferences, or chronic sample contamination. The sample itself may generate real or apparent bias caused by a matrix effect or variation in physical properties such as particle size. A bias DQI is a quantitative indicator of the magnitude of systematic error resulting from these effects. Bias can be in the positive (high) or negative (low) direction from the true value and is usually unknown in magnitude.

Accuracy is a measure of the overall agreement of a measurement to a known value. In a limiting case where random errors are very tightly controlled, then bias dominates the overall accuracy. In general, however, both precision and bias contribute to accuracy. A measurement result even with zero bias may not be accurate if the measurement process is not precise. Figure 2-2 demonstrates how different combinations of precision and bias can contribute to accuracy. The true mean, indicated by the dashed line, is 10. Diamond symbols indicate the observed measurements. Part A of Figure 2-2 shows the least desirable situation of significant bias and low precision. The accuracy of these data is poor. Part B of Figure 2-2 shows the case of low bias and low precision. Any one of the measurements is not accurate, but an average of all the measurements would be. Part C of Figure 2-2 shows data with significant bias and high precision. The accuracy of these data are poor. In this case, random errors are well controlled, but a systematic error limits the accuracy of each individual measurement. Part D of Figure 2-2 shows the case of low bias and high precision. Each individual measurement is accurate.

Representativeness, as defined by the American Society for Quality and published in the American National Standards Institute (ANSI) document, ANSI/ASQC E4-1994, *Specifications and Guidelines for Quality Systems for Environmental Data Collection and Environmental Technology Programs* (ANSI/ASQC, 1994), is "The measure of the degree to which data accurately and precisely represent a characteristic of a population, parameter variations at a sampling point, a process condition, or an environmental condition." Developing a clear understanding of the "population" that is the subject of an experiment or investigation is the key to assessing representativeness. The characteristics of the population include the subject's identity or class (e.g., the particular property that needs to be measured), the spatial distribution of the property, and in some cases, the temporal characteristics of the property. Representativeness is usually considered a qualitative term that does not lend itself to being measured by a DQI. However, there are quantitative measures that may be applied and are discussed in Section 4.1.

Comparability is the qualitative term that expresses the measure of confidence that two or more data sets can contribute to a common analysis. Before pooling data, the comparability of data sets generated at different times or different organizations must be evaluated in order to establish whether two data sets can be considered equivalent in regard to the measurement of a

specific variable or groups of variables. In a laboratory analysis, comparability is influenced by the analytical method used, laboratory performance, holding times, and sample handling procedures such as sieving or mixing. Comparability is also related to representativeness. Data sets that are representative of two different populations are generally not comparable with respect to pooling. However, if other aspects of the data collection protocol are comparable, these data may be useful in testing statistical hypotheses.

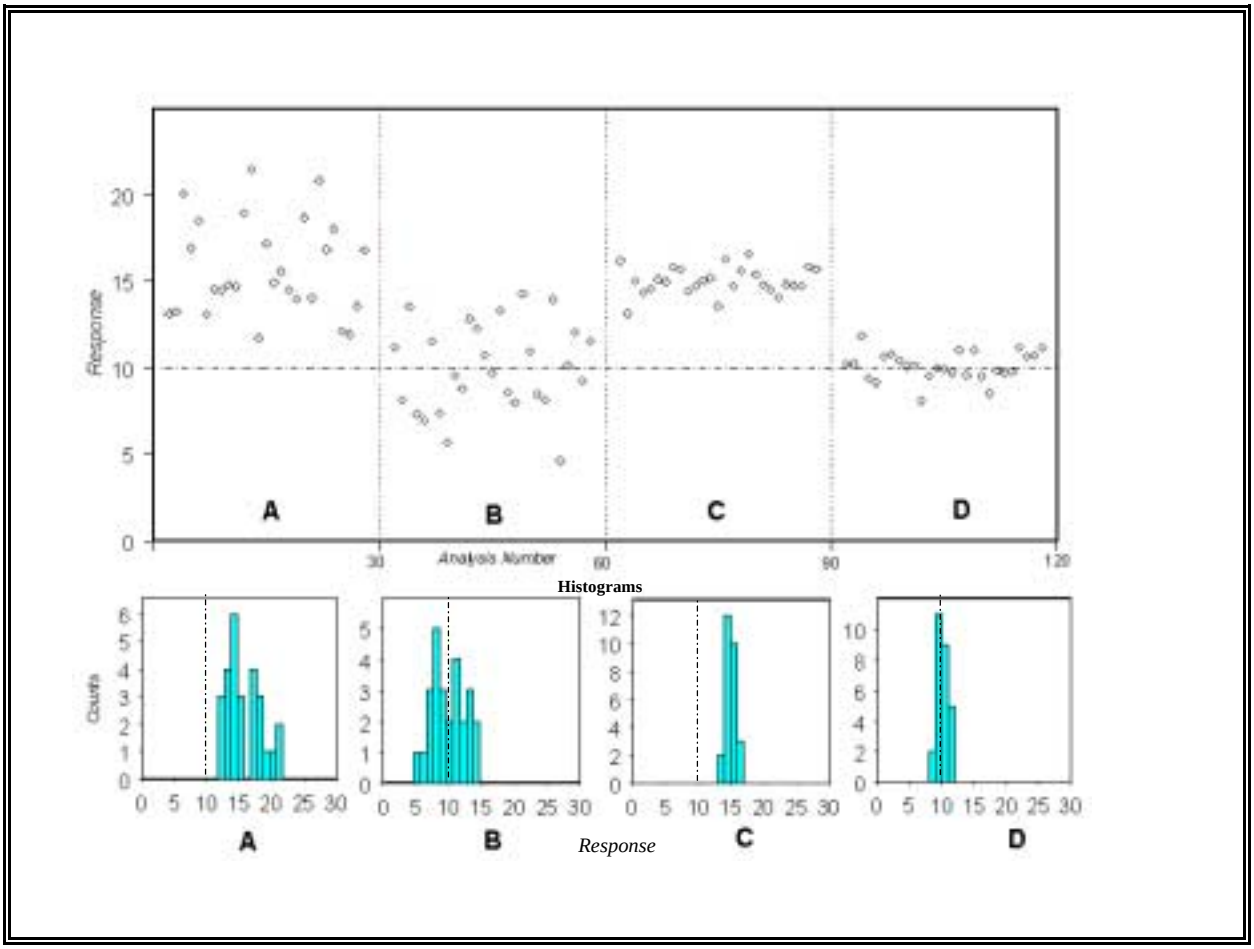


Figure 2-2. Influence of Bias and Precision on Accuracy

Completeness is a measure of the amount of valid data obtained from a measurement system, expressed as a percentage of the number of valid measurements that should have been collected (i.e., measurements that were planned to be collected). Low completeness can have serious effects on statistical analyses because of loss of statistical power in the design. The DQI for completeness is typically expressed as a percentage. For example, 70% completeness implies 30% of the planned measurements were lost or found invalid.

Sensitivity is the capability of a method or instrument to discriminate between measurement responses representing different levels of the variable of interest. The term "detection limit" is closely related to sensitivity and is often used synonymously. In practical applications, sensitivity is the minimum attribute level that a method or instrument can measure with a desired level of precision. Sensitivity is often a crucial aspect of environmental investigations that must make comparisons to particular action levels or standards. Details regarding specifications and use of sensitivity DQIs are provided in Section 3.5.

2.2 FRAMEWORK FOR DECOMPOSING COMPONENTS OF VARIABILITY AND BIAS

Before discussing details of DQI calculations and uses, a discussion of the framework used in this document for examining error sources is necessary. This document uses the model depicted in Figure 2-3 as the basis for discussing DQIs that correspond to specific components of total study error. Total study error is a statistical measure of the uncertainty in a metric, such as a site mean concentration, caused by the combination of all error sources in the study design. The term "error sources" refers to any factors that build uncertainty into a measured value. Generally speaking, these error sources are the result of natural variability in the sampled media and inherent imprecision in the measurement process. Mistakes resulting in contaminated or lost samples are outside the statistical definition of total study error. The framework for examining study errors will also serve as a context for clarifying the relationship among total study error, project DQOs, and MQOs [also referred to as measurement performance criteria in *Guidance on QA Project Plans (EPA QA/G-5)* (U.S. EPA, 1998a) and *Requirements for QA Project Plans (EPA QA/R-5)* (U.S. EPA, 2001b)].

At an elementary level, the breakdown of total study error starts by dividing total error into two main compartments: error related to obtaining measurements *within* sampling units, and error associated with variability and bias *between* sampling units. A sampling unit is the portion of the physical environment from which one or more samples may be taken, resulting in measurement(s) appropriate for an intended use. Typically, sampling units comprise the members of the population that may be selected for sampling, such as individual objects (people, trees, etc.) or specific areas or volumes of environmental media for which a measurement is desired. As will be discussed further in the next section, the identification of sampling units for a given project may depend both on the physical characteristics of the media being studied, and on statistical issues pertinent to the study. Sampling units, as defined here, can also be referred to as “units of analysis.”

The contribution to total error from within-unit sources is largely a function of several factors, including the way sampling units are defined for the study at hand, the inherent variability of the characteristic of interest within the unit, the difficulties associated with obtaining a specimen from the unit, and the error associated with the measurement process itself.

2.3 APPROACHES TO DEFINING SAMPLING UNITS

The most common application of sampling units is to assume that the unit is equivalent to the physical sample taken (e.g., a 4-inch core, a specified volume of air, a simple grab sample, or a composite of multiple grabs over some specified area or time). Under some circumstances, alternative definitions of sampling units may be useful.

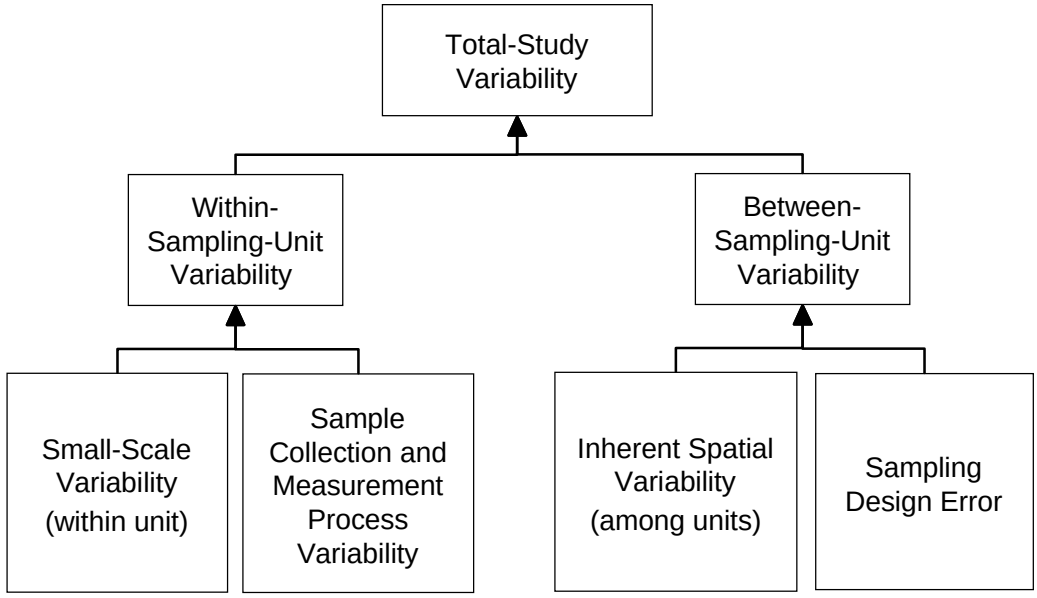


Figure 2-3. Simple Total Study Error Model

Sampling Units Larger than the Size of the Actual Physical Sample

- Limitations of the physical sampling approach may necessitate definition of a sampling unit larger than the size of the actual sample. A desire for composite sampling over a small area (or volume), a need to collect multiple specimens in order to obtain enough of the media of interest to conduct the requested analyses, or sampling equipment for which precise siting cannot be accomplished, are examples of situations that might benefit from a broader definition of a sampling unit.

Sampling Unit Equal to a Multiple of Separate Physical Samples

- When collection of field duplicate samples, or collocated samples, is desired, the sampling unit may be defined as just the size necessary for collection of a minimum of two separate samples.

Sampling Unit Defined Uniquely

- Special definition of sampling units will be necessary in some cases, for instance: where the support of the physical specimen is too small to provide a meaningful representation of the population; where there is a desire to control the precision of measurements over a small area, volume, or time period (possibly for contouring of the concentrations); where significant auto-correlation of nearby sample locations is expected based on the conceptual model, and available data indicate that nearby samples may violate the assumption of independence; or when the physical specimen is not sufficient to be an adequate surrogate for a population unit as required by classical statistical sampling protocols.

Once the sampling unit definition is complete, within-unit sources of error can be examined. The within-unit error sources, as shown in Figure 2-3, consist of small-scale variability and measurement error. Small-scale variability refers to the natural variability of the measured property on the scale of the sampling unit. Measurement error comprises all sources of error involved in the activity of obtaining a value for the property of interest. The measurement process may include the process of obtaining one or more physical specimens from within the sampling unit, combining samples to represent the unit, subsequent subsampling in the laboratory, and finally conducting a physical measurement (such as a chemical analysis for contaminants of concern).

Defining sampling units as some area, volume, or time period larger than the size of an individual physical specimen has the potential benefit of clarifying whether collocated (or simultaneous) samples should be treated as additional field samples, or as samples that represent the same unit (replicates). Collocated samples have historically been considered by many members of the environmental quality assurance community as the best type of sample to estimate total measurement system (specimen collection and measurement) error [e.g., *A Rationale for the Assessment of Errors in the Sampling of Soils* (EPA, 1990)]. However, statisticians have argued that some of the variability observed between collocated samples is the result of small-scale variability and the related problem of adequately representing this variability by obtaining one or more samples from within the sampling unit. Clarifying that collocated samples provide an estimate of within-unit error removes the confusion. If we assume that between-sampling-unit variance components dominate total study error (which is often the case in environmental studies), the within-unit spatial variance component can often be left out of practically all equations to determine the number of samples to collect, including all sample size formulae presented in *Guidance on Sampling Design* (U.S. EPA, 2000e). However, should within-unit variance components be found to contribute significantly to total study variance, then statistical sample designs should consider trade-offs between obtaining replicates within units, or collecting samples from additional units. In addition, when evaluating alternative measurement methods, within-unit replicates of cheaper screening methods may be a cost-effective alternative to more expensive and precise fixed laboratory methods, assuming that other characteristics of the method are adequate for the intended use (e.g., sensitivity and bias).

2.4 ESTABLISHING MQOs IN THE CONTEXT OF DQOs

One way to employ DQIs is as a means of specifying data quality criteria (MQOs) which, if achieved, will provide an indication that the resulting data are expected to meet the DQOs. Used in this way, DQIs provide a metric against which the performance of a program can be measured during the implementation and/or assessment time frames. The process of establishing MQOs is intertwined with the process of designing a study, which is beyond the scope of this guidance document. However, general descriptions are provided in the following text as well as in Example 2-1.

During the design phase, the type and number of samples required to achieve DQOs, and the way in which these samples should be optimally allocated across space and time, are developed. *Guidance on Sampling Design* (U.S. EPA, 2000e) provides guidance on alternative approaches to solving this problem. In most cases, to determine the number of samples needed to achieve a certain statistical performance goal, a relevant estimate of total study precision is required. This estimate generally comes from historical data. In many cases, historical data and information are available, but were collected with older, or different, sampling and/or analytical methods than are currently being considered. To determine the potential impact of using newer methodologies on total study precision, it is useful to see how much these within-unit error sources contribute to the total study error. Assuming the newer methods are more precise, and the design team feels it is safe to assume that these improvements will be realized, specific MQOs can be established, and these improvements can be factored into the design equation.

A practical approach to accomplish this is to first determine the performance required for the DQI, based on an analysis of the relative impact of a specific error source or quality attribute on total study design. Next, determine whether the performance that can be expected from the particular instrument, method, or laboratory is as good as or better than required, and adopt these defaults as the MQOs. If defaults are not adequate, then MQOs should be specified, and a new agreement reached with a potential vendor (or a new instrument or method selected) to meet or exceed these more stringent requirements. If aimed at the analytical laboratory, this may take the form of a special analytical request or a performance-based method contract. If aimed at the use of an instrument or method, it may take the form of a new standard operating procedure designed to ensure that the MQO is achieved (for example, standard operating procedures for replicating x-ray fluorescence (XRF) measurements to achieve a precision requirement).

Example 2-1. DQOs and Associated MQOs

Step 1: State the Problem

Fisherman's Bay, Maine, has been a favorite location for clamming by the local population (both private and commercial) for decades. Local health officials have routinely monitored the bay as part of the comprehensive marine shellfish toxins monitoring program to ensure that red tide has not affected this area and that the clams are safe to eat. Recently this monitoring has been augmented in Fisherman's Bay to include the evaluation of organic and inorganic constituent levels in tissues. So far, there has not been a need to restrict clamming in this area; however, a local citizen's group has expressed concern that the monitoring program does not include any clams from near the local industrial complex, which due to it's easy access and large mud flats is known to be a productive area for clamming. In particular, polychlorinated biphenyls (PCBs) and mercury are contaminants of potential concern, and have been the target for soil remediation activities at the industrial complex. A study is needed to determine if the concentration of PCBs or mercury in shellfish in the bay near the industrial complex are elevated with respect to reference locations in the bay. If so, a human health risk assessment should be conducted, and if necessary a fishing advisory posted. In addition, if clams in this area are unacceptably contaminated, the need for sediment or additional soil remediation should be evaluated.

Step 2: Identify the Decision

The following study question was identified:

Do concentrations of total PCBs or mercury in clams harvested from the vicinity of the industrial complex in Fisherman's bay exceed those in clams from other (ambient) locations in Fisherman's Bay?

Step 3: Identify Inputs to the Decision

Inputs include:

- results from the analysis of clam tissue from clams representative of areas surrounding the industrial complex and from areas representative of ambient conditions in Fisherman's Bay and
- analysis of tissues should, at a minimum, be for total PCBs (estimated as the sum of 18 specific congeners listed by NOAA) and mercury.

Step 4: Define the Study Boundaries

The study should address surficial sediments (top 6 inches) from a 10.5-acre area that surrounds the industrial facility, and from three reference locations selected based on historical measurements of sediments, similarity in terms of grain size and TOC, lack of any point sources and based on agreements with the regulatory authorities. The concentration of contaminants in tissues is not expected to be seasonally influenced, therefore measuring contaminants in tissues once a year, during the spring sampling campaign, should be adequate.

Step 5: Develop a Decision Rule

If the mean concentration of total PCBs or Hg in clam tissue from the area surrounding the industrial complex is significantly greater than the mean concentration in clams collected from the ambient locations, then a human health risk assessment will be conducted to evaluate risk due to ingestion of clams from the industrialized area.

Example 2-1. DQOs and Associated MQOs

Step 6: Specify Tolerable Limits on Decision Errors

The probability of failing to determine that clam tissue concentrations in the industrialized area are greater than ambient, when in "truth" they are elevated by 50%, will be limited to 5%, and the probability of incorrectly determining they are the same, to 10%. Failure to properly determine that clams in the industrialized area are more contaminated would result in a failure to evaluate or communicate increased risk due to clamming in this area. Improperly determining that clams in the industrialized area are elevated over ambient would result in falsely alarming the public, unnecessary expenditures related to conducting risk assessments, and unnecessary reduction in the harvest and associated lost of revenue to commercial clamming operations.

Step 7: Optimize the Design for Obtaining Data

Based on sample size calculations utilizing variance estimates from existing clam studies, nine composite samples (comprised of 8 to 10 clams each to generate the required amount of tissue to conduct the analyses) from the industrialized area, and nine composites from the three ambient locations (three from each) are required to achieved the specified error constraints.

MQOs Associated with Data Collection

Measurement quality objectives for analytical detection limits and other indicators of data quality were developed in support of this study.

- Sensitivity: MDLs for compounds of interest are determined annually according to 40 CFR 136 (Appendix B) by spiking clean, low-lipid tissue with all parameters of interest and then processing them according to the CFR methods. Calibration curves include a concentration close to the expected MDL, and this low point of the calibration curve is used as the reporting limit (RL), unless it is determined that it is below the MDL, or outside the linear range of the method. The RL for Hg based on the proposed analytical method is approximately 0.00095 mg/kg. The risk based concentration of potential interest is in the range of 1.2 to 60 mg/kg, and historical tissue in relatively clean areas has been measured at .025 mg/kg. Total PCBs are calculated from a sum of congeners, therefore no RL for total PCBs is available, however the risk range of potential concern is from 45 to 7,800 μg/kg. Historical total PCBs have been observed at 1.2 to 220 μg/kg, and detection limits for congeners using the proposed method have not been a problem.
- Precision: PCBs have historically been more variable than Hg, and were used to drive the design for this study. Total standard deviation of the historical measurements was 142 μg/kg. Based on an evaluation of laboratory splits, field duplicates, and the normal field samples, 80% of the total standard deviation was associated with between-unit (spatial) variability. Default laboratory precision limits are deemed adequate and adopted as MQOs for the laboratory contribution of total variability.
- Bias: Laboratory defaults for acceptable recovery were adopted as MQOs for this study. For PCBs, 75-125% recovery using EPA method 608 (gas chromatography/electron capture detector), and for Hg recovery of 85-115% using EPA method 245.5 (cold vapor atomic absorption).
- Representativeness: To ensure that samples of clams are representative of each area, each composite sample will be generated from clams collected from fixed points along randomly oriented transects located in each area of interest.
- Completeness: A minimum of 8 of the 9 required composites from the industrialized area, and 8 of the 9 required composites from the reference areas are required to achieve the error constraints, if historical estimates of variability are confirmed. Every attempt should be made to achieve 100% completeness.
- Comparability: To maximize the utility of the data, the same laboratory, sampling and analytical methods as are used in the general monitoring program will be employed in this study. The chain-of-

CHAPTER 3

DQIs RELATED TO ENVIRONMENTAL MEASUREMENTS

3.1 MEASUREMENT PROCESS

In environmental studies, measurement is the process of obtaining a quantitative value describing a chemical or physical property of an individual sampling unit or specimen collected from this unit. Measurement may involve direct field measurements using survey instruments or collection and handling of physical samples followed by analysis in a fixed or mobile laboratory. Choices that the technical team makes regarding sample acquisition, sample handling, preparation, and analysis can influence the quality attributes of the resulting data. Sources of measurement variability include:

- within-unit, small-scale variability (influenced by the nature and distribution of the characteristic of interest within the sampling unit and the media assayed),
- physical sample acquisition protocol (sample collection tools, procedures such as compositing),
- sample handling and transport,
- variability associated with the measurement assay (influenced by the capabilities and management of the analytical resources),
- homogenization and subsampling procedures,
- sample preparation and extraction procedures,
- subsampling extract for analysis, and
- analytical determination.

Figures 3-1 and 3-2 present an example of a measurement process that involves several individual steps. Each step has the potential to introduce variability or bias that might influence the quality of project data. Estimates of the various components of overall study variability and the relative contribution of measurement (within unit) precision are important inputs to the statistical design process. A QA Project Plan describes the processes used to estimate and monitor the magnitude of at least the more important potential measurement error sources. The quality planning activity for measurements simultaneously establishes MQOs appropriate for the project and data use and defines the required DQIs. Very often, the data sources for DQIs are derived from samples inserted in the sample stream by field or laboratory personnel on a frequency specified in the QA Project Plan. The bottom tiers of Figures 3-1 and 3-2 suggest the types of QC samples that might be identified as the source of the underlying data for precision (Figure 3-1) and bias (Figure 3-2) DQI calculations.

Results from routine quality control samples often can be used to construct useful DQIs; however, some thought must be given to what support such data provide to the quality goals set out in the QA Project Plan. One use of routine quality control results that is always valid is assessment of the laboratory's internal operations; for example, the routine analysis of calibration

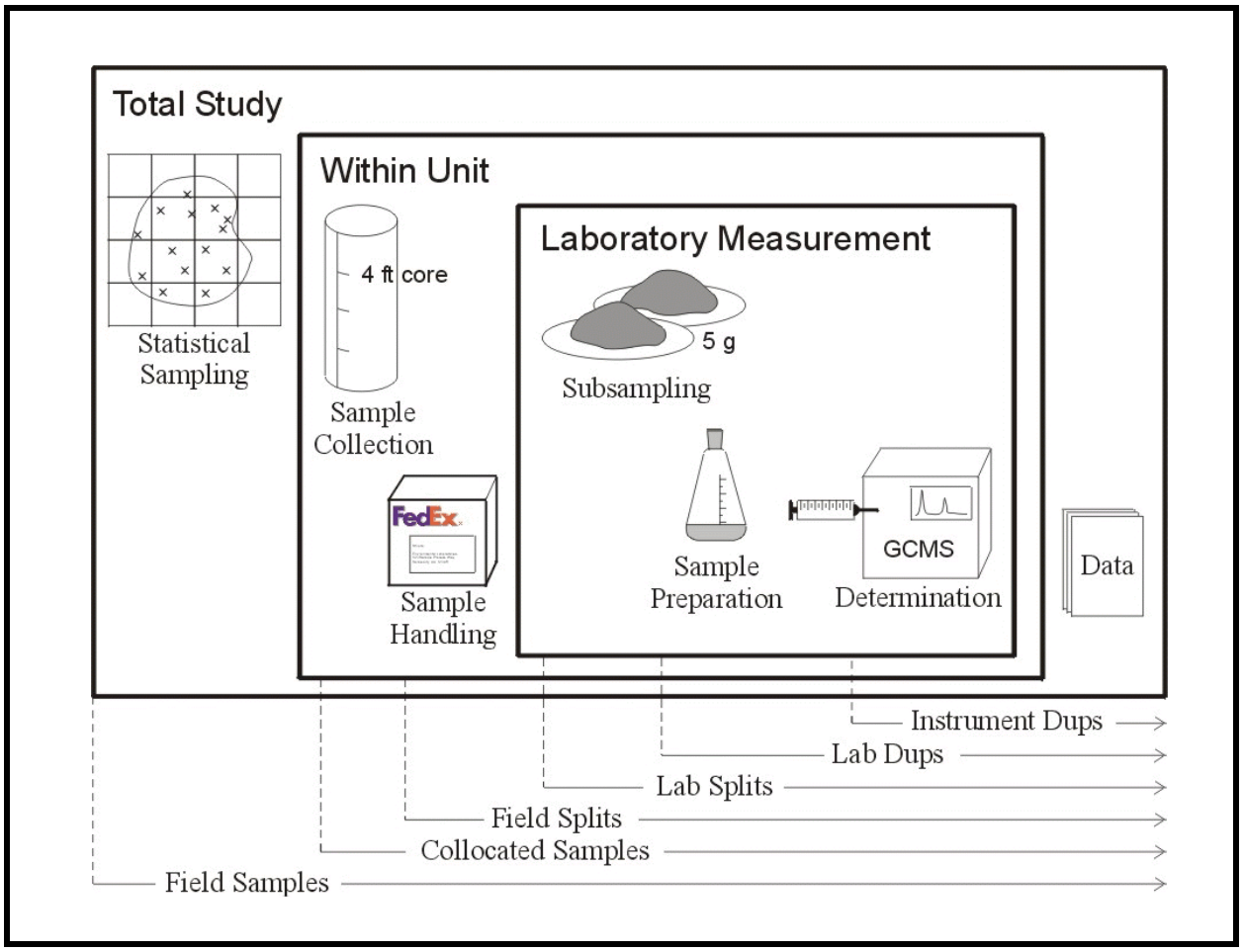


Figure 3-1. Total Sampling and Measurement Process Denoting the Use of QA Samples to Measure Components of Total Study Precision

blanks to test for contamination or a problem in system calibration. In this case, a DQI could be just the actual blank result, expressed in appropriate units, with an MQO equal to the laboratory's internal acceptance criterion. A poor result from a routine calibration blank should result in an immediate corrective action by the laboratory. While it is important to verify the laboratory is monitoring and controlling the quality of their internal operations, results from samples like a calibration blank have little use in project planning from the DQO perspective. The same can be said for overall analytical calibrations. A properly calibrated analytical system is always required and expected for analysis of project samples. Calibration data verify internal quality control, which in turn is assumed to take place and is therefore not directly discussed during the DQO planning process. Instead, the quality attributes of greatest utility from a DQO planning and statistical design perspective include estimates of the overall (total) study variability and an understanding of the relative contribution of significant components of this total.

3.2 MEASUREMENT PRECISION

Precision is a data quality attribute upon which statistically designed sampling is based. Therefore, DQIs for precision are among the most important quality indicators in an environmental study. Precision is estimated by using some form of replication followed by calculation of a DQI based on the replicate measurements. Precision DQIs provide a measure of agreement among replicate analyses. MQOs for precision set tolerable limits of imprecision for a study.

3.2.1 Common Indicators of Precision

When the purpose of the QC check is to simply provide inputs to routine data verification procedures, thus providing a baseline level of confidence in laboratory performance, the MQOs associated with the laboratory's performance can simply be a listing of the laboratory's internal control limits. The underlying assumption is that in the design process, the performance specifications of the selected laboratory (i.e., precision, bias, sensitivity) were evaluated and judged adequate to achieve the project DQOs. However, a more thoughtful analysis of MQOs, driven by the DQO process, should always be considered before listing laboratory "default" limits in a QA Project Plan. This is especially important when the project does not require levels as stringent as the laboratory's default limits to be adequate for the intended use. For example, if a default practical quantitation limit (PQL) is well below levels of concern, a project-specific requirement should be stated in the QA Project Plan. In that way, data do not get flagged during a data validation when they are adequate to support the intended use.

The most common precision DQIs are the summary statistics range (R), for duplicate measurements, and standard deviation(s), for multiple measurements. The range of duplicates is calculated as the absolute difference between the two values; that is:

$$R = \left| x_1 - x_2 \right|.$$

The standard deviation is calculated by the statistical formula:

$$s = \sqrt{\frac{\sum \left(X_i - \overline{X} \right)^2}{n-1}}$$

where n is the number of replicate analyses. The square of the standard deviation, s^2, is an important statistic called the variance. The variance is often used to describe the spread of a data set.

Because the precision of environmental measurement systems is often a function of the property being measured (e.g., as concentration increases, standard deviation generally also

increases), precision indicators expressed relative to the average of the replicate analyses can be useful. Two relative precision indicators, relative range (RR) (for duplicates) and relative standard deviation (RSD), find frequent application as DQIs. The RSD is also known as the coefficient of variation (CV). The RR and RSD formulae are simply:

$$RR = \frac{\left|x_1 - x_2\right|}{\overline{x}}$$

$$RSD = \frac{s}{\overline{x}}$$

where $\overline{x}$ denotes the arithmetic mean of the *n* samples. Both RR and RSD can be expressed as a percentage by multiplying the indicator by 100%. The relative range expressed as a percentage is often called the relative percent difference (RPD).

3.2.2 Effect of Concentration on Measurement Precision

An evaluation of the precision DQI as a function of concentration (or other measured property) will usually lead to one of three conclusions:

1. standard deviation (or range) is independent of concentration (i.e., constant);

2. standard deviation (or range) is directly proportional to concentration so that the coefficient of variation (or relative range) is constant; or

3. both standard deviation (or range) and coefficient of variation (or relative range) vary with concentration.

Creating plots of s (or R) or CV (or RR) versus concentration can help clarify which indicator is best suited for a particular data stream. Figure 3-2 demonstrates the different relationships between precision indicators and the underlying population characteristic (e.g., concentration). The indicator with the most simple behavior should be selected for use; i.e., for case (1) the standard deviation or range is simplest to work with, whereas for case (2), the coefficient of variation or relative range is simplest. If the relationship of precision to concentration falls into case (3), the relationship might have to be modeled using regression analysis or other statistical means to estimate how standard deviation (or range) varies with concentration. If either the standard deviation (or range), or the coefficient of variation (or relative range) is approximately constant in the range of interest to the decision-maker, then case (1) or (2) essentially holds.

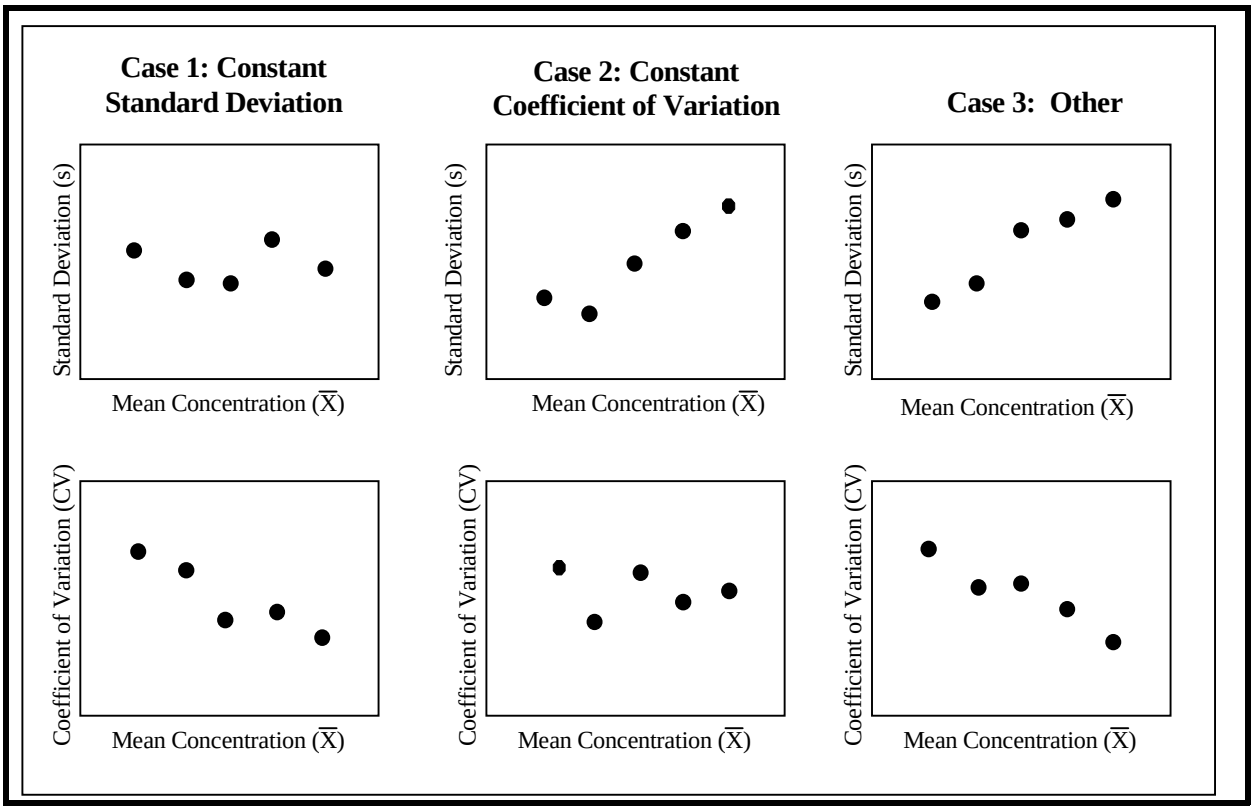


Figure 3-2. Relationship Between Precision and Concentration

3.2.3 Components of Within-Unit Precision

The overall precision within a measurement process can be broken down to different levels by use of replicate analyses. Figure 3-3 lists the types of physical samples and the measurement activities that may be monitored by the replicated analyses. Each type of replicate generally incorporates several sources of error. If the sampling unit is defined as an area large enough to obtain multiple specimens, results from "collocated" samples provide an overall estimate of the within-unit precision. This includes the sample acquisition and handling processes, measurement process, and inherent small-scale variability. If the sampling unit is defined as the size of the physical specimen, then field or laboratory splits can be used to estimate the within-unit error term. Figure 3-3 illustrates the additive nature of components of variance, showing how different QA samples provide information on different parts of the whole.

The breakdown of precision components shown in Figure 3-3 might seem to imply the measurements occur at the same time and the analytical work is performed by a single laboratory. However, in larger projects, two or more laboratories may be involved with measurements conducted over a significant length of time. Replicate samples can readily provide indicators of the effects on precision from these interlaboratory factors as well. For example, portions of field

split samples may be sent to two or more laboratories in order to generate an indicator for interlaboratory precision.

Figure 3-3 uses a laboratory-based analyses sequence to demonstrate how QC samples support calculation of DQIs. Characterizing and monitoring the performance of field measurements should be performed in an analogous fashion in order to assess precision of repeated measurements taken by the same field instrument on the same sample or sample location. On larger projects, variation among different instruments used for the same type of measurement, and variation among different field technicians collecting field measurements using the same instrument and the same sample, may need to be monitored with a DQI.

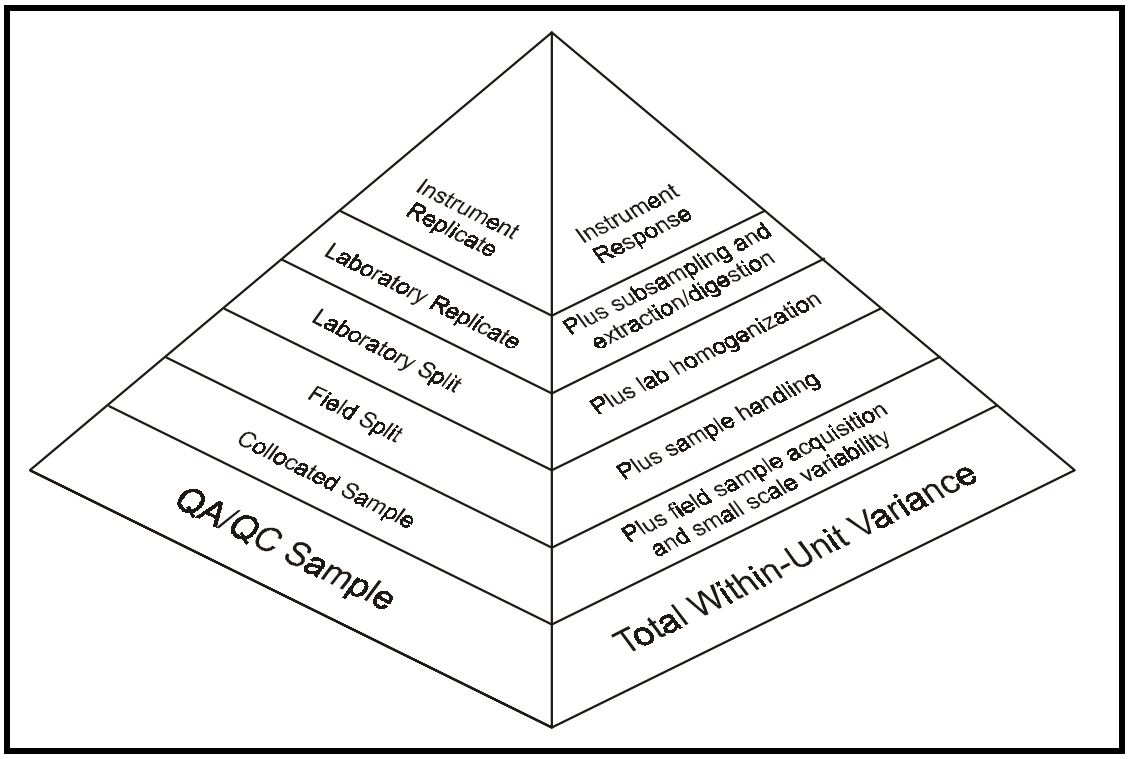


Figure 3-3. Total Within-Unit Precision Pyramid

In the event that calibration or other routine quality control activities are not monitored as needed, the impact on data quality cannot be known, but may be severe and threaten the success of any project. Therefore, appropriate laboratory quality control must not be simply "assumed." All expected routine laboratory QC operations must still be documented in the project QA Project Plan.

3.2.4 Establish Measurement Quality Objectives for Precision

Total study variability is an important input used in determining the sample size required to achieve some desired performance, as specified through the DQO process. MQOs on specific measurement components should reflect the requirements of total study error associated with the measurement process so that the DQOs may be met. In order to establish the MQOs, the sensitivity of overall study error to the precision of various subcomponents of the total (e.g., the analytical component of measurement error) should be examined. A simplistic approach for examining the relative contribution of different components of total study variability begins from the assumption that total variance can be broken into individual, additive, components:

$$\sigma_T^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \ldots\ldots$$

Using the framework for decomposition of errors shown in Figures 3-1 and 3-4, the sum of errors in an environmental investigation might be

$$\sigma_T^2 = \sigma_b^2 + \sigma_w^2 = \sigma_b^2 + \sigma_s^2 + \sigma_m^2$$

where the subscripts' definitions are: T = total study, w = within-sampling-unit, b = between-sampling unit, s = small-scale, and m = measurement.

Working with variances can be difficult conceptually because the terms are squared. A visualization tool using the more intuitive standard deviation statistic may be used for examining the additive relationship between precision components. The visualization tool takes advantage of Pythagorus's Theorem concerning the sides of a right triangle (the square of the hypotenuse is equal to the sum of the squares of the other two sides). Figure 3-4 is a graphical representation of the most basic division of total study variability into the within-unit (measurement and small-scale) and between-unit (field or spatial) variabilities that was previously introduced in Figure 2-3. Note that in this model the lengths of the sides of the triangle are directly proportional to the standard deviation, or precision, of the different components. The total standard deviation is obtained from the results of the actual randomly located field samples while the estimate of within-unit precision is best estimated using collocated samples. The between-unit variability is inferred from the other two sides of the triangle. The dimensions in Figure 3-4 represent the relative scales often encountered in environmental sampling, that is, between-unit variability generally dominates total study variability.

The precision of the sample collection and measurement process could be broken down even further as suggested by Figures 3-6 and 3-7. These measurement components can be difficult to characterize independently because a QC sample chosen to help with the characterization must be carried through the entire measurement process in order to obtain the result, thus integrating effects of all measurement components after the point the QC sample is created.

Mathematically, the variances of the additive components of total variability can be expressed using the following formulae. These formulae depict the mathematical equivalent of Figure 3-5.

$$s_T^2 = s_b^2 + s_w^2$$
$$= s_b^2 + s_s^2 + s_m^2$$
$$= s_b^2 + s_v^2 + s_h^2 + s_p^2 + s_i^2$$

where T = total; s = within-sample; h = homogenization; b = between-unit; m = measurement method; p = sample preparation; w = within-unit; v = inherent within-sample variability; and i = instrument.

Example 3-1. Relative Contribution of Components of Total Variance

The table below presents a summary of the variance calculated based on an arsenic study where a full range of samples were collected in sufficient numbers to generate estimates of several important components of total study variance. In this example, as in many environmental studies, the total study variance is dominated by spatial variability (between sampling unit variability).

Sample	Components of Variability Captured	Estimated Standard Deviation
Instrument replicate	Instrument response	0.046
Laboratory replicate	Instrument response plus subsampling and extraction/digestion	0.11
Laboratory split	Instrument response, subsampling and extraction/digestion, plus laboratory homogenization	0.12
Field split	Instrument response, subsampling and extraction/digestion, laboratory homogenization plus sample handling	0.12
Collocated samples	Instrument response, subsampling and extraction/digestion, laboratory homogenization, sample handling, plus field sample acquisition and small scale variability	0.15

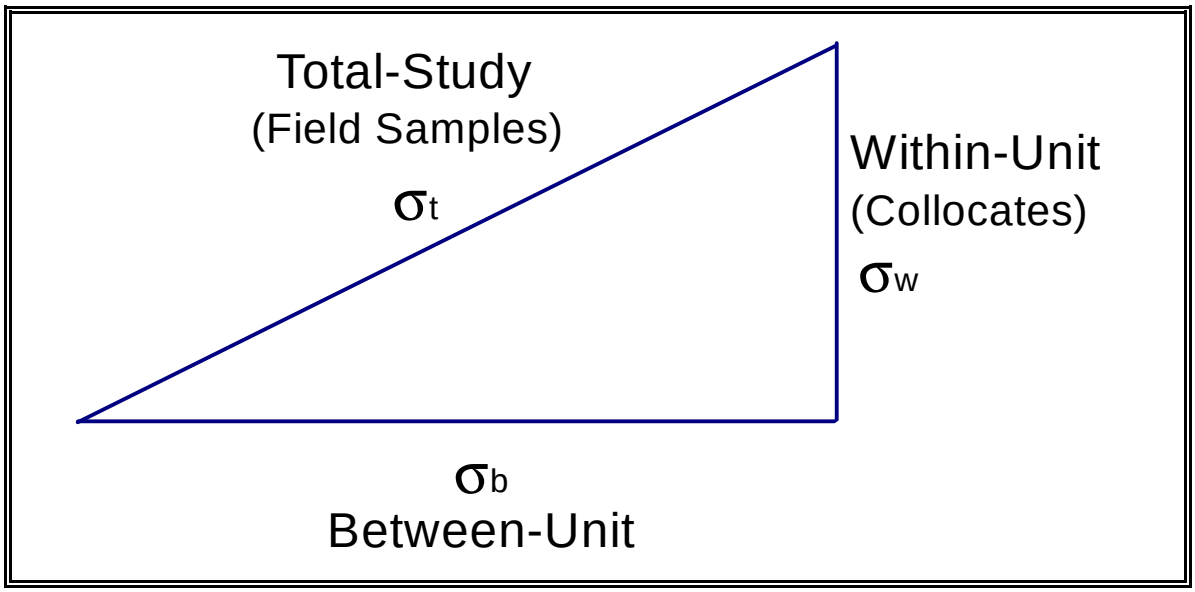


Figure 3-4. Use of Right-Triangle to Represent the Influence of Variance Components on Total Study Variance

Example 3-2. Evaluating Between-Unit Variance

Data collected from across a petroleum spill site were analyzed with appropriate QC samples as dictated by the QA Project Plan. The within-unit standard deviation was calculated from a set of collocated samples to be 7.5 parts per million (ppm). The total standard deviation of the site-wide results was 20.5 ppm. Based on the relationship between within-unit, between-unit, and total study error, the between-unit error component can be estimated to be:

$$s_b = \sqrt{\left((20.5)^2 - (7.5)^2\right)} = 19.1 \text{ ppm}$$

An MQO established for this study was that not more than 80% of the total study error (expressed using standard deviation) be attributable to between-unit differences. The between-unit precision in the observed data accounts for over 90% of total study error. This result raises concern over the adequacy of the design which was based on assumptions that were not borne out by the data. Careful appraisal of data adequacy for decision-making may be necessary.

Note that the theoretical standard deviation, σ, has been replaced by the sample standard deviation, *s*, because the calculation based on data provides only an estimate (*s*) of the standard deviation of the population.

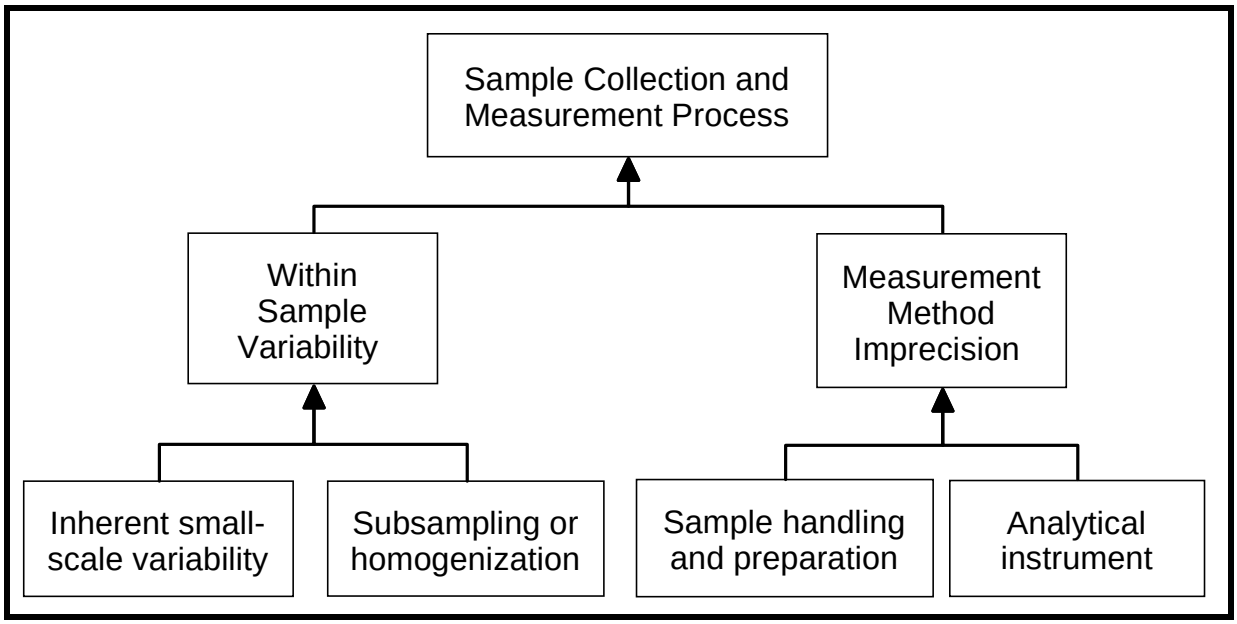


Figure 3-5. Components of Within-Unit Variance

Figure 3-6 demonstrates how the precision of the measurement and small-scale variability affects the within-unit variability. The numbers in the correspond to hypothetical standard deviations of each component. In a similar way, intra-laboratory and inter-laboratory precision can be analyzed.

The within-unit standard deviation is estimated using results from collocated samples as discussed above. The measurement precision is likewise estimated from results of laboratory replicates. The small-scale variability must be inferred from the other two sides of the triangle. Taking the process a step further, Figure 3-6 shows how sample preparation variability can be inferred from laboratory replicate and instrument replicate results.

Understanding the relative importance of components of total study variance is key to establishing precision MQOs based on DQOs. The MQOs may be stated in terms of the total sample error, or more specifically in terms of various components of that error.

It is important to note that the equations presented for adding components of variance do not always make sense with actual data. Because an investigator can only estimate the variance through collection of a small sample from the population, these formulas may not hold for the

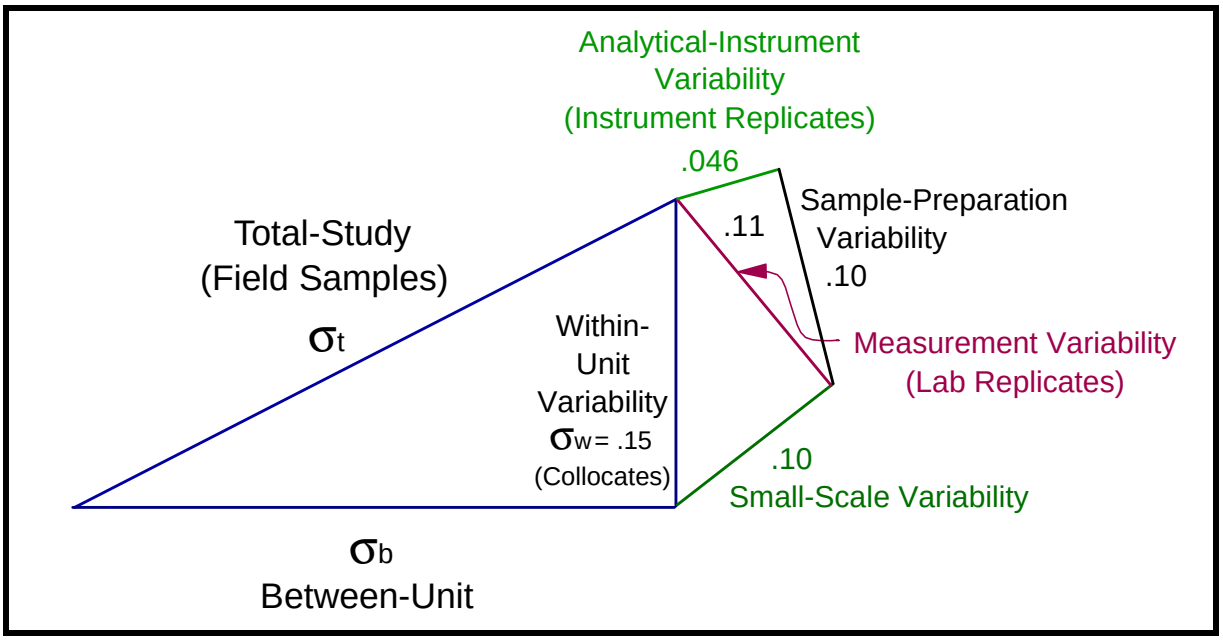


Figure 3-6. Components of Within-Unit Variance Depicted Using Right Triangles

estimates. In theory, the components of variance may be broken out as shown. In practice, it is possible for the measurement variance to exceed the within-unit variance. For example, suppose that an estimate of small-scale variability was calculated to be a negative number. This would indicate that this component of variability was extremely small, and could be regarded as being essentially zero. In addition, the additive variance formula is based on the assumption that within-unit variability is independent of between-unit variability. This is an assumption that may be violated in environmental problems, and the investigator should maintain an awareness of that situation. If between-and within-unit variability are strongly correlated, a more complex model for the composition of the total study error may be required. A model with terms for interaction effects between error components would account for the lack of independence. When data for various error components is correlated, a statistician should assist in development of an appropriate model.

3.2.5 Collection of Replicate Samples or Measurements Within the Sampling Unit as Required to Achieve Within-Unit MQOs for Precision

The use of replicates at various stages of analysis can be of great help in reducing specific components of variance. Replication should therefore be considered as an option when determining how to achieve MQOs for precision. If sampling units are defined as something larger than the size of the physical sample taken, any two samples within the sampling unit would be considered a replicate, and would be useful in reducing or assessing within-unit variance. If multiple readings or analyses are performed, and the average used, the variance of that component is reduced dramatically, and the contribution of variability to the total within-unit variance is reduced. It follows that the best tactic to reduce total within-unit variance is to

replicate those components that contribute the greatest proportion to this variance term. Likewise, before looking at ways to reduce within-unit variance, the relative contribution of this component to the total study variance should be assessed. Usually greater gains in the reduction of the total study variance are made by sampling additional sampling units to address between-unit variability, rather than focusing on the within-unit components.

Example 3-3. Utilizing QC Data to Evaluate Variance Components

These data come from a site where thorium-228 was of potential concern. Based on a statistical sampling design, eleven samples were collected from the site. The sampling plan also called for a variety of QC samples, including field splits and laboratory replicates. Several more studies were to be conducted on this and similar sites following this small study, so it was important to the project team to estimate various components of study error. If they found that one component had a significant impact on the total study error, it might be worth investigating methods for reducing that error. Specifically, if the measurement error was great, it might be worth doing more expensive and precise measurements, where possible. If the between-unit variability is great, that is an inherent part of the system which can not be changed, but further sampling designs could be planned to accommodate that understanding.

Location ID	Sample ID	Value	Unit	Depth	Sample Type
z37p-1860	99-037-7510	1.646	pCi/g	0-3"	
z37p-1860	99-037-7510	1.359	pCi/g	0-3"	laboratory replicate
z37p-1860	99-037-7511	0.5814	pCi/g	3-6"	
z37p-1860	99-037-7511a	1.384	pCi/g	3-6"	field split
z37p-1860	99-037-7512	1.785	pCi/g	6-12"	
z37p-1861	99-037-7513	0.7812	pCi/g	0-3"	
z37p-1861	99-037-7515	1.141	pCi/g	6-12"	
z37p-1861	99-037-7515	1.63	pCi/g	6-12"	laboratory replicate
z37p-1862	99-037-7516	1.443	pCi/g	0-3"	
z37p-1862	99-037-7518	1.234	pCi/g	6-12"	
z37p-1862	99-037-7518a	1.179	pCi/g	6-12"	field split
z37p-2568	99-037-7275	0.9872	pCi/g	0-3"	
z37p-2568	99-037-7275a	1.0346	pCi/g	0-3"	field split
z37p-2568	99-037-7276	1.207	pCi/g	3-6"	
z37p-2569	99-037-7279	1.127	pCi/g	3-6"	
z37p-2569	99-037-7279	1.0782	pCi/g	3-6"	laboratory
z37p-2569	99-037-7280	1.946	pCi/g	6-12"	

Example 3-3 (continued)

Within-unit variability may be calculated from the field split data. The following provides one estimate of precision based on this data:

$$s_w^2 = \frac{\sum (x_n - x_{nsplit})^2}{2k}$$, where k = the number of locations with field splits.

$$s_w^2 = \frac{(0.5814 - 1.384)^2 + (1.234 - 1.179)^2 + (0.9872 - 1.0346)^2}{2 \times 3} = 0.108$$

Based on all samples, the total study variance is estimated to be:

$$s_t^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1} = 0.907$$

From these two pieces of information, and the triangular relationship between total study error, within-unit variability, and between-unit variability, it is possible to estimate:

$$s_b^2 = s_t^2 - s_w^2 = 0.907 - 0.108 = 0.799$$

The proportion of the total study variability comprised of within-unit variability is estimated as:

$$\frac{\sqrt{s_w^2}}{\sqrt{s_t^2}} = \frac{0.329}{0.952} = 0.35$$

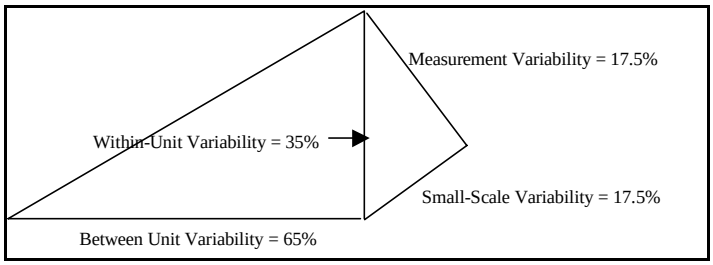
It is now apparent that 35% of the variability observed in this data set is from within-unit sources. As the data set also contains laboratory replicates, the within-unit variability can be broken into measurement error and small-scale variability. In this study, sampling units were defined as the size required to obtain a single sample. The field method for obtaining samples gathered more of the media than necessary for analysis, and the field splits were taken from homogenized single grab samples. Thus, the small-scale variability estimate represents the inherent variability within a single grab sample.

$$s_m^2 = \frac{(1.646 - 1.359)^2 + (1.141 - 1.63)^2 + (1.127 - 1.0782)^2}{2 \times 3} = 0.054$$

Finally, it can now be estimated that the small-scale variability is:

Example 3-3 (continued)

Interestingly, for this data set, the small-scale variability and measurement error are both equal contributors to the within-unit variability, both contributing 17.5% of total study error.



3.3 MEASUREMENT BIAS (ACCURACY)

Bias in a measurement system may arise from a number of sources ranging from chronic sample contamination issues to unreliable calibration. Bias is normally estimated from testing the measurement system result against a specimen with known properties. The most common DQIs for bias are derived from the results of QC samples such as spiked samples, standard reference materials, and various kinds of blanks in the sample stream. Table 3-1 lists some common QC samples and the components of bias they are intended to measure.

Table 3-1. QC Samples for Deriving Bias Indicators

Sample Type	Indicator For
Blank spike	Instrument contamination or malfunction, calibration shift
Matrix spike	*plus* effectiveness of sample extraction/digestion procedures
Reference material	Same as matrix spike, but more representative of overall performance when material is similar to matrix examined in the study
Calibration blank	Instrument contamination, calibration shift
Preparation blank	*plus* laboratory contamination
Field blank (equipment/trip)	*plus* field, transportation, and storage contamination

The difference between the measured and expected result is a DQI for bias. For spikes and reference materials, a bias indicator is most conveniently expressed as a fractional or percent

comparison of the measured result to the expected result. Relative bias is also a common indicator, which indicates both the magnitude and direction (positive or negative) of the bias. Expected results are based on the known properties of the QC sample.

The expected result is usually the true value derived from a standard or a theoretical value as indicated by the project's QA Project Plan.

$$relative\ bias = \frac{measured\ result - expected\ result}{expected\ result}$$

Relative bias is sometimes used to describe bias in terms analogous to relative precision. For applications of bias with percent recovery of a known contaminant, a commonly used measure of bias is percent recovery which is simply:

$$percent\ recovery = \frac{measured\ result}{expected\ result} \text{ X } 100\%$$

Matrix spikes can also be used to determine percent recovery as follows:

$$percent\ recovery = \left(\frac{x_s - x_u}{k}\right) \times 100\%,$$

where x_s = measured value of the spiked sample
x_u = measured value of the unspiked sample, and
k = known amount of spike in the sample.

A completely unbiased result thus has recovery of 1 (or 100%) and recovery may be greater or less than 1 (100%) depending on whether the result is higher or lower than the known quantity. For blank samples, the actual magnitude of the result is the DQI because the "known" quantity should be zero for a blank.

Minor short-term problems that clearly might bias an individual measurement but do not create a systematic bias are best reflected in precision DQIs. For example, if an analyst makes a small error on a dilution as part of the sample preparation step, that mistake will bias the individual result; however, this sort of error is part of the overall precision in the measurement process. Because the analysis of QC samples listed in Table 3-1 must include the entire measurement step, the precision of the measurement process is also imposed on results intended for bias DQIs. Except in the case of large relative bias, it can be difficult to assess if recovery DQIs reflect systematic bias or normal measurement variability. For this reason, bias DQIs primarily serve a quality control function during implementation of a project plan. The sampling

and analysis phases of a project should be developed on the assumption of an unbiased measurement system. During the planning phase, procedures put in place to minimize the potential for bias, and the use of bias DQIs to verify the procedures, should have the desired effect of keeping bias to a minimum.

Example 3-4. Estimating Relative Bias

Bias is generally evaluated against some "known" value. For example, reference materials, matrix spikes, or blanks are evaluated to determine the bias of the analytic method. In this case, the investigator has done substantial work with lead analyses conducted by inductively coupled plasma/mass spectrometer (ICP/MS), and is willing to accept that analytical method as sufficiently unbiased and precise for his purposes. He is now curious about the bias of field screening using an XRF device relative to the ICP/MS results. Because there is an abundance of data (n=55) from the same locations (the XRF reading was taken, then the sample removed and sent to the laboratory for ICP/MS analysis), the mean of the recoveries from each pair of samples can be calculated. This provides an estimate of the overall bias between methods.

Location ID	XRF In-Situ Result (ppm)	Fixed Laboratory Analysis (ppm)	Relative Bias
1	59	133	-0.56
2	13	23.4	-0.44
3	18	60.1	-0.70
4	1.0	15.4	-0.94
5	144	101	0.43
6	37	37.2	-0.01
7	1.7	25.8	-0.93
8	0.4	33.2	-0.99
9	470	106	3.43
10	6.2	14.4	-0.57
11	4.0	26	-0.85
12	245	198	0.24
13	9.5	11.5	-0.17
etc.			

The mean estimated recovery of XRF in-situ analysis for lead based on all 55 samples was determined to be 93%. This is equivalent to a negative bias of 7% from the expected (or true) values.

Example 3-5. Evaluating Recovery

A reference material sample of known concentration was analyzed three times to determine the bias, if any, of laboratory analyses associated with a batch of groundwater samples. The level of aroclor 1264 in the reference material was 1.25 ppm. The results of the analyses on the reference material were 0.87, 0.91, and 1.14 ppm. The average of these readings is 0.97. An estimate of the bias in aroclor 1264 results may be given by the equation:

$$recovery = \frac{0.97}{1.25} \times 100\% = 77.6\% \quad .$$

The MQO for bias for this study was plus or minus 20% recovery (80 to 120%). The 78% recovery should be considered in assessing of data adequacy, and should result in appropriate

Example 3-6. Impact of Non-Responses on Study Bias

A national survey on wildlife encounters was sent to a random sample of 1000 townships. 60% of the queried counties responded to a question on the number of wildlife encounters reported in their county in the preceding 12-month period. The average reported number of annual encounters was 100. Is it reasonable to assume that the average number of wildlife encounters reported by those counties is representative of the nation? Let's consider some possibilities:

If the actual average number of wildlife encounters in the non-responding counties was also equal to 100, then the bias = 0% and the correct national estimate is 100 per county.

If the actual average number of wildlife encounters in the non-responding counties was equal to 90, then the bias = 4% and the correct national estimate is 96 per county.

If the actual average number of wildlife encounters in the non-responding counties was equal to 70, then the bias = 14% and the correct national estimate is 88 per county.

If the actual average number of wildlife encounters in the non-responding counties was equal to 50, then the bias = 25% and the correct national estimate is 80 per county.

The impact of the non-respondents may be great if there is any bias in the "selection" process. That is, if there is a difference between the characteristics of the respondents and non-respondents related to the topic of interest, then an estimate based on incomplete data may be biased.

Example 3-7. Use of Internal Standards and Surrogates to Assess Bias

Problems with analysis and bias can also be indicated by close examination of internal standards and surrogates, common QC components of organic analysis. In this example, a laboratory was analyzing groundwater samples for purgable organics following Solid Waste Disposal Act Method 524.2, "Measurement of Purgable Organic Compounds in Water by Capillary Column Gas Chromatograph/Mass Spectrometer (GC/MS)" (U.S. EPA, 1992). Following an initial calibration check sample and laboratory reagent blank, eight water samples were analyzed. This analytical method requires the addition of an internal standard (fluorobenzene) and two surrogates (1,2 dichlorobenzene d-4, and bromofluorobenzene). These chemicals would not be expected to be found in groundwater. The role of surrogates is to mimic the behavior of target analytes. Thus they monitor the efficiency of preparation/extraction but do not evaluate bias directly (they are not the actual target compounds).

QC Data Summary Table for Organic Analyses

			Concentration in Sample								
Analyte	PQL= RL (μg/L)	MDL, μg/L (25 mL sample)	1	2	3	4	5	6	7	8	8b
Acetone	0.550	0.122	1.300	0.880	<0.12	1.10	1.90	0.650	0.790	1.10	1.50
Benzene	0.210	0.051	0.275	0.289	<0.05	0.27	0.28	0.24	0.210	0.12J	0.22
2-Butanone	0.753	0.144	<0.14	0.880	<0.14	0.891	0.799	0.793	0.794	.035J	0.911
MTBE	0.322	0.061	1.80	2.20	0.350	2.10	1.10	0.750	0.700	0.600	0.69
o-Xylene	0.375	0.068	1.100	0.880	<0.07	1.04	0.88	0.585	0.585	0.55	0.71
Internal Standard (% rec.)			104	108	101	98	100	88	85	78	107
1,2 dichlorobenzene (SA, % Rec.)			100	99	103	100	104	85	83	80	105
Bromofluorobenzene (SA, % Rec.)			101	103	107	101	96	90	87	81	100

MDL = method detection limit; PQL = practical quantitation limit; MTBE = Methyl tertiary butyl ether; SA = surrogate analyte; J = detected between the MDL and PQL

The results of nine analytical runs are shown. The samples are ordered by the sequence in which they were analyzed. When the data are reconstructed in this fashion, and recovery of the internal standard and surrogates is examined through the work shift, a problem can be seen starting with sample 6. This trend is also depicted in the following graph.

DQIs that measure bias are also valuable tools for ensuring comparability of data. Completeness of a data set also has the potential to impact bias. If a data set is incomplete, any systematic trend to the missing data may cause bias in estimates of population parameters based on the data set.

3.4 COMPARABILITY

Comparability is the qualitative term that expresses the confidence that two data sets can contribute to common interpretation and analysis. Quantitative measures of comparability are also available involving statistical tests that measure the similarity or difference between two or more data sets. Comparability must be carefully evaluated in order to establish whether two data sets can be considered equivalent in regard to the measurement of a specific variable or groups of variables (U.S. EPA, 1998a, U.S. EPA, 1997).

Comparability is a very important qualitative data indicator for analytical assessment, and is critical when considering the combination of data sets with the same analytes. The assessment of this DQI determines if analytical results being reported are equivalent to data obtained from similar analyses. Only comparable data sets can be readily combined.

Data may be considered comparable, and decisions based on the combined data sets may be appropriate for some problems, while the same data are not sufficiently comparable for other problems. As with any decision about the usability of data, it is important to consider the decision that the data are meant to support. Separate determinations of the comparability of data sets may be necessary for each decision the data are used to support.

A number of issues to consider prior to data analysis may illuminate the incomparability of two data sets. These issues are usually described and compared qualitatively.

3.4.1 Measures of Comparability

Table 3-2 presents nine indicators of comparability, and some questions that should be considered related to each. The presence of these nine indicators enhances the comparability of distinct data sets:

These characteristics vary in importance depending on the final use of the data. The closer two data sets are with regard to these characteristics, the more appropriate it will be to compare or combine them. It is also possible that large differences between characteristics may be of only minor importance, depending on the decision that is to be made from the data. Each procedure and method used must be fully described, validated, and performed by competent practitioners, and performance should be evaluated against a reference. This standard applies to both sample collection procedures and methods used in the field as well as procedures and methods used in the analytical laboratories.

Table 3-2. Indicators of Comparability

Indication of Comparability	Related Questions
Samples within data sets should be selected in a similar manner	*Sample design*: Were the samples selected in a similar manner? Are they equally representative of the population of interest? If samples in one data set were selected using a judgmental sample design, and another data set is based on a statistical design, then combining these data may not be appropriate for some uses.
Data should be temporally and spatially consistent	*Sample collection dates*: Were samples collected in the same sampling event? Are there temporal factors such as seasonality or holding times that could directly affect interpretation of the data?
	Sample location: Were the samples taken from the same area? Are they representative of the same population spatially? If they are from different areas, how are they expected to be similar? How are they expected to differ?
	Matrix: Were the samples from the same matrix? This relates to how the samples were collected, location of the samples, and when the samples were collected. If matrices are different, are they expected to be related in some way?
Data sets should contain the same set of variables of interest	*Variables of Interest*: Which variables are of interest and are necessary for grouping or analyzing the data? Were these variables reported for all data sets? For example, particle size, total organic carbon, or percent moisture may be useful for determining if the data are comparable, but these variables may not always be reported.
Units in which these variables were measured should be convertible to a common metric	*Units*: Units should be reported for all data sets. Are the units all convertible to a common metric? For example, some results may be reported in wet weight and some in dry weight, which are not directly comparable without additional information.
Field collection methods should be similar	*Field methods*: What instrument was used and which procedure was followed? Were single or composited samples collected?
	Sample handling: Some samples require special handling such as preservatives or special containers. Differences in sample handling may cause variations in the results, which may affect comparability. Were the samples filtered or unfiltered? Are there chain-of-custody forms available for all samples?

Table 3-2. Indicators of Comparability

Indication of Comparability	Related Questions
Similar sample preparation methods should be used	*Laboratory*: Was the same laboratory used for all analyses? The use of routine methods and procedures simplify the issues of comparability because the same standards should be met. In addition, this will increase confidence in the comparability of methods used.
	Sample preparation: Was the same sample preparation used for all samples? If not, are the sample preparation methods comparable? For example, sample preparation may be a total digestion compared with a partial digestion, which would not result in directly comparable results.
Similar procedures and quality assurance should be used to collect and analyze samples for all data sets	*Analytical method*: Was the same analytical method used for all samples? If not, are any of the analytical methods comparable? The use of routine methods simplifies the determination of comparability because all laboratories used the same standardized procedures and reporting parameters. However, when reviewing the analytical methods, consideration must also be given to options that may be available within the method. Although the analytical method may be the same, options such as matrix or concentration level will affect results reported.
	Analytical method options: If the analytical methods are comparable, were the same options within each method chosen? The options available within each method must also be checked because the same analytical method using different options may produce very different results.
Measuring devices used for both data sets should have approximately similar detection levels	*Detection or quantitation level*: Are non-detects generally reported at the same level? Are the detection or quantitation levels acceptable for use in decision making? Combining data sets having different detection or quantitation levels leads to difficulties in analytical interpretations.
	Quality control of data entry, storage, transfer, and retrieval: Were results reported into the database in a consistent manner? Have all data sets been checked for completeness?
Rules for excluding certain types of observations should be similar for all data sets	*Qualification and/or validation of data*: What criteria were used to qualify or validate the data? If criteria were not consistent across data sets, the same qualifications may have different meanings. What QA and QC information is available from the laboratories?

Example 3-8. Assessing Comparability of Two Data Sets

Data were collected by a government agency to determine if contamination existed at a coastal site. The city that owns the land adjacent to the government facility also sampled this area. The primary concern for both parties, if there were contamination, was to determine if the source were upstream on city land or at the government facility. Most of the city's data were from upstream of the site of interest, with only eight samples on the coastline. The government agency's data were from along the coastline and into the inter- and sub-tidal zones. Combining these data sets and assessing the trend in concentrations over space would provide some insight into the location of the source of contamination, if one exists.

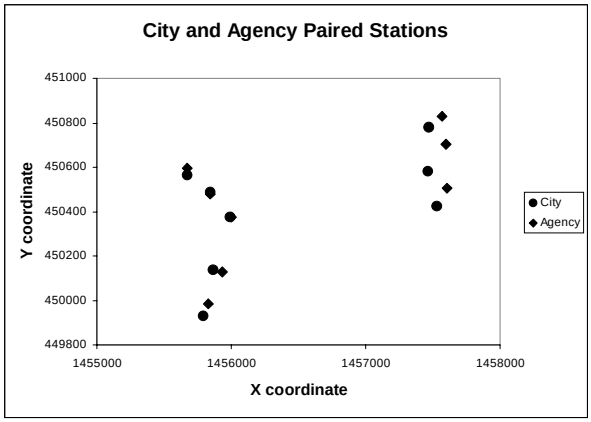
	Data Attribute	Data Set #1: City of White Springs	Data Set #2: Government Agency	Remarks
Field	Sample collection method	unknown	spade and scoop	
	Matrix	sediment	sediment	
	Sample handling	unknown	according to approved standard operating procedure	
	Sampling event	10/1997	9/1998	
Analytical	Sample preparation	unknown	Method 3050: Acid Digestion of Sediments, Sludges, and Soils	
	Analytical method	SW-846	SW-846	
	Analytical method option	- -	- -	
	Detection level	680 - 1240	1 - 3.5	
	Units	ppb	ppm	can be converted to match
	Fields of interest that were reported	all desired fields except QC data and sample collection method	all desired fields	locations in different scales: conversion and matching of sample locations was possible
	Criteria for exclusion of samples	none	Rosner's test for outliers	no data were excluded based on these rules

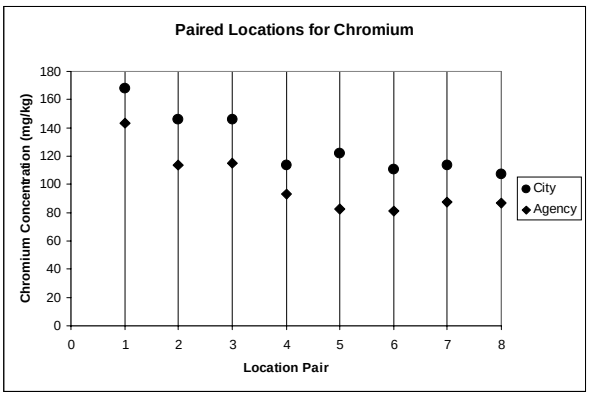
Based on the information presented in this data comparability summary table, it is difficult to ascertain whether or not these two data sets are comparable. All of the necessary information to determine if the sample collection and analysis methods were comparable is not available. However, we do know that both sets of samples were analyzed according to SW-846 methods, and the data are reported with similar detection limits.

An empirical comparison of results from those locations that were most closely matched was conducted. The eight coastal locations from which the city collected samples are closely collocated with sample locations from which the government agency collected samples. These eight locations are presented below.

Example 3-8 (continued)

DATA OWNER	LOCATION ID	RESULT	QUALIFIER	UNITS
City	1	168		mg/Kg
City	2	146		mg/Kg
City	3	146		mg/Kg
City	4	114		mg/Kg
City	5	122		mg/Kg
City	6	111		mg/Kg
City	7	114		mg/Kg
City	8	107		mg/Kg
Agency	1	143		mg/Kg
Agency	2	114		mg/Kg
Agency	3	115		mg/Kg
Agency	4	93.4		mg/Kg
Agency	5	82.9		mg/Kg
Agency	6	81.5		mg/Kg
Agency	7	87.3		mg/Kg
Agency	8	86.5		mg/Kg





Although sufficient information is contained in the table to begin to ascertain the comparability of these data sets, a visual display of the data clarifies an important point. At every one of the eight similar sampling locations, the chromium concentration reported by the city is greater than the value reported by the agency. Although the year between studies could be responsible for some of that difference, the conceptual model does not support that change over time. It would be prudent in this case to halt any joint use of these data until some of the uncertainty over sample collection and sample preparation methods can be removed. There is a clear difference between the results from these two distinct data sets that is not fully explicable with the available information, and comparability should be strongly questioned.

3.4.2 Comparability and Combining Data Sets

Comparability is very important when conducting meta-analysis, which combines the results of numerous studies to identify commonalities that are then hypothesized to hold over a

range of experimental conditions. Meta-analysis can be very misleading if the studies being evaluated are not truly comparable. Without proper consideration of comparability, the findings of the meta-analysis may be the result of an artifact of methodological differences among the studies rather than differences in experimentally-controlled conditions. The use of expert opinion to classify the importance of differences in characteristics among data sets is invaluable.

Example 3-9. Importance of Maintaining Meta-Data to Assess Comparability

For longitudinal studies, such as annual national surveys of water quality or river usage, it is important to maintain consistent meta-data. Meta-data, such as records on the date of data collection, the name of the person who provided the data, temperature and precipitation data for the year, and other descriptive information are vital to these surveys. It is well known that questions that are rephrased between versions of a study may elicit different responses. Similarly, changes in other aspects of the study will introduce additional variability in the results. When the county utilities manager who has responded to the survey for many years retires, the new respondent may provide different answers based on knowledge, experience, and interpretation.

The key to comparability is consistency of approach, which applies to both the field portion of the sampling and the laboratory analysis of the samples. For studies where new data will be collected and decisions will be based only on this set of data, comparability can be managed by establishing clear specifications for sampling, analysis, quality control, and data reporting. The data should be comparable if the project is carried out according to the plan even if several different parties are involved.

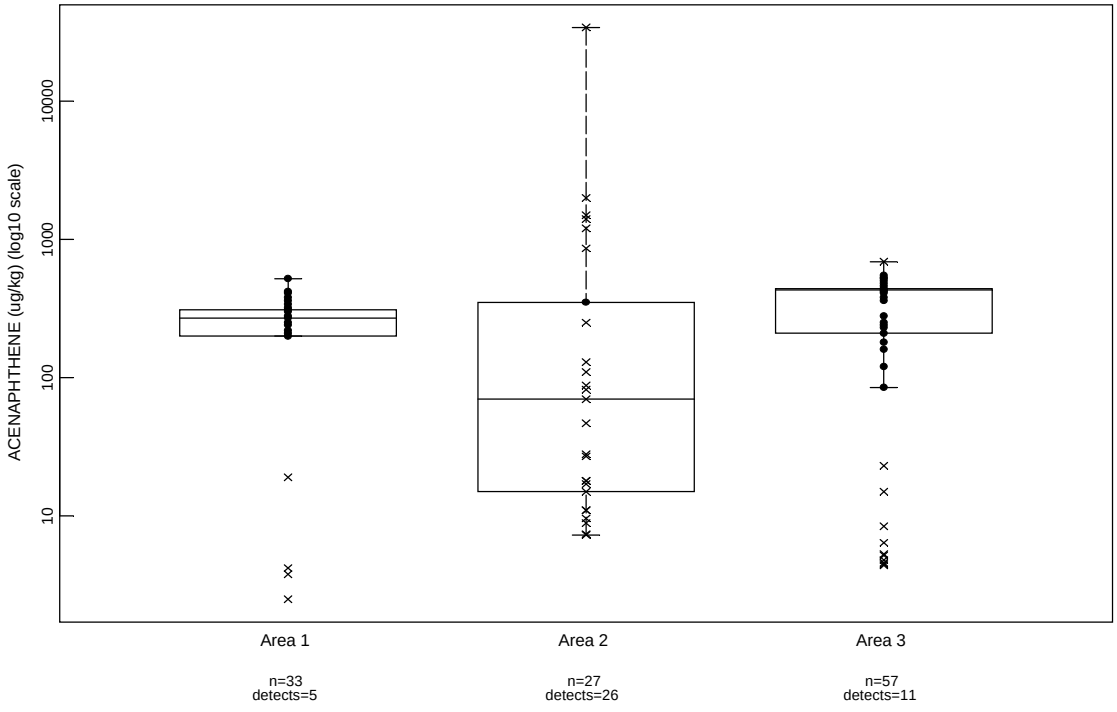
For studies where new data will be collected, and it is known that historical data will be combined with the newly collected data, additional precautions should be taken. The planning phase should include consideration of historical data and how the results were produced. For the field portion of the design, it is important that studies be planned such that they do not provide spatially disparate representations of the population of interest. It is also necessary to consider temporal aspects such as time-lag between studies if the site being investigated is part of a dynamic system.

With respect to the laboratory analysis portion of the design, it is not necessarily in the best interest of the project to use the same analytical method, sample preparation, etc., for the new samples. This may be due to new analytical methods or procedures that will provide more sensitivity for the analysis or because the method used for the historical data was not the optimal method that could have been used at the time the samples were collected. In cases where new samples will be collected, a chemist should be consulted to evaluate the comparability of the methods used to analyze the historical samples and the methods proposed to analyze the new samples. If these methods are not directly comparable, one may consider running a subset of the new samples by both methods for comparability purposes (Gilbert, 1987, Chapter 9: Double

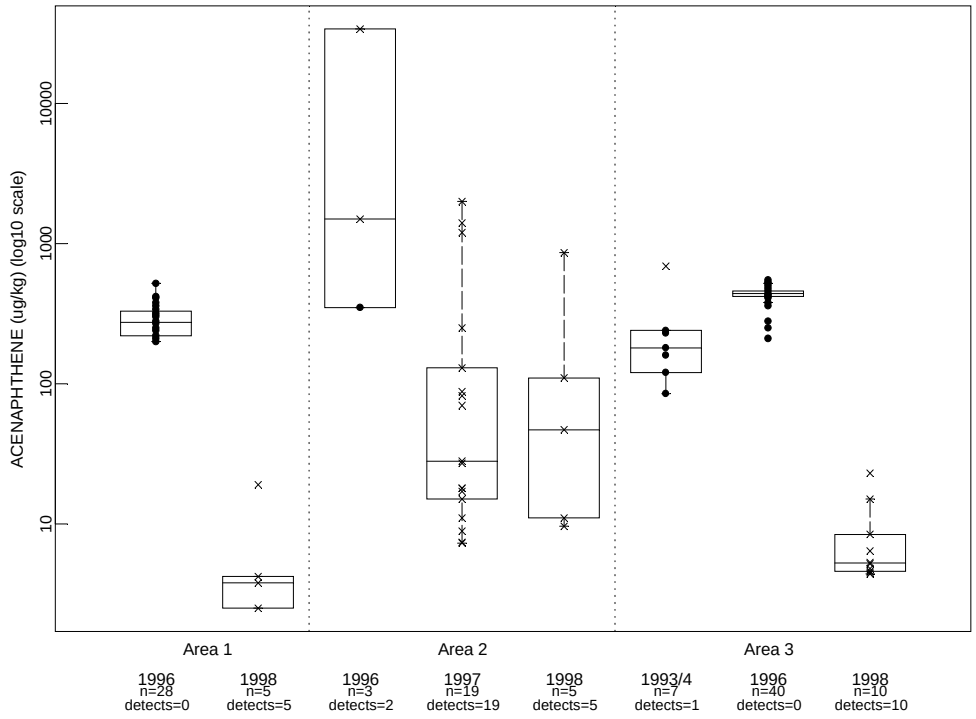
Sampling). This type of planning will allow for overlap between the new and historical data where the results can be directly compared, thus providing information about the differences between the data sets so the data can be used appropriately. Although the two data sets may not be directly comparable, the overlap of data will provide information about how the data are related, and how all of the data may be used to support a decision.

Example 3-10. Use of Side-by-Side Box Plots

Samples from three related areas have been collected over the six-year period from 1993 through 1998. The data for each area are presented in the box plots below. Below each box plot is the sample size, and number of detections. The box represents the bulk of the data (specifically, the 25th, 50th, and 75th percentiles of the data), while each data point is also shown individually. Detects are represented as "x"; non-detects as "•".
A quick glance at these box plots shows that the median concentrations for sites 1 and 3 are greater, and the range less, than for site 2. This presentation shows something about the comparability of the overall data for the sites, but misses looking at comparability of the separate sampling campaigns. The separate sampling campaigns are presented by year in the following figure.



Example 3-10 (continued)



3.4.3 Statistical Comparability

If data sets are deemed qualitatively comparable, then many methods are available for determining if they are statistically comparable. Statistical measures should not be applied until qualitative comparability is achieved, because purely quantitative comparisons can be misleading. For example, a data set on the number of pilot whales in the Pacific Ocean and a data set on the number of trees in Jefferson County, Colorado, that are infected with mistletoe may have similar means and variances, but obviously should not be combined or considered comparable under any circumstances.

Exploratory data analysis using simple graphics is often an effective tool to support comparability assessments. Exploratory data analysis may involve visualization techniques such as box plots, scatter plots, histograms, or quantile plots (Cleveland, 1993). For example, one can

see if the data appear to be from the same distribution by looking at box plots. If the data do not appear to be from the same distribution, perhaps a scatter plot of each analyte compared with a normalizing factor, such as total organic carbon or percent moisture, would show a relationship between the data sets. When appropriate documentation cannot be recovered, certain elements, such as reporting/censoring limits, may be deduced from plots of the distribution of values in a data set. This information should always be used in conjunction with all other information that is available before determining comparability of the data sets.

Statistical comparisons of means, variances, distributions, or other parameters provide quantifiable information on data comparability. Many techniques such as *t*-tests and similar non-parametric tests (Wilcoxon Rank Sum test, Kruskal-Wallis test) are appropriate for comparing characteristics of environmental data distributions (U.S. EPA, 1996; Gilbert, 1987). Occasionally, statistical comparisons are the only tools available for determining comparability. When that situation arises, great care should be taken in interpreting the results.

Example 3-11. When Comparability Cannot be Assessed

Data collected on emissions in a western city were compiled from three databases for analysis. The study question was whether there were substantial differences in emission rates across makes of automobile. Each of the three data sets contained information on the automobile manufacturer, year of vehicle, and emissions rating for hundreds or thousands of vehicles. A variety of additional data were available in one or two data sets, but no other data fields were complete and available in all three data sets. Qualitative measures of comparability were not assessable, because of the lack of meta-data on methods or DQIs, so statistical comparisons of comparability were conducted. The Wilcoxon test for similarity of distributions showed that while two data sets had comparable emissions rating results, the other data set was statistically different. Researchers could not determine whether the difference was caused by temporal, measurement, or other explainable phenomenon, so the third data set was not used further in the city's analysis.

3.5 SENSITIVITY

Sensitivity generally refers to the capability of a method or instrument to discriminate between small differences in analyte concentration. Both the precision of the instrument and the slope of the calibration curve limit sensitivity. Chemists typically define sensitivity as the slope of the calibration curve at the concentration of interest (Skoog, 1985). If two methods have equal precision, the one having a steeper calibration curve will be the more sensitive. Sensitivity can also be evaluated from the standard deviation of replicate analyses at any concentration level, or can be evaluated from the confidence bound on a calibration curve. It represents the minimum difference in two samples that can be distinguished with a defined confidence (Taylor, 1987).

The sensitivity indicators of primary interest to EPA are indicators that relate to limits of detection. In determining the detection limit, the focus is on the concentration that can be distinguished from the noise of the method. A number of indicators related to sensitivity will be discussed. Differences exist between IDLs, MDLs, RLs, and PQLs. In addition, a number of related indicators such as the LOD, LOQ, minimum level, and decision level have been proposed in the literature and will be discussed.

3.5.1 Detection Limit Concept

The detection limit (DL) is a concept concerning the capability of an analytical method to distinguish samples that do not contain a specific analyte from samples that contain low concentrations of the analyte. It is important to note that many analytical methods produce non-zero signals even when a target analyte is not present. The DL is generally considered to be the minimum true concentration of an analyte producing a non-zero signal that can be distinguished from the signals generated when no concentration of the analyte is present, with an adequate degree of certainty. In other words, the DL is a value above which the probability of finding an analyte is present, when in truth it is not, (generally referred to as a false positive) is adequately small.

DLs vary by analyte and by matrix, and often vary among laboratories, so it is critical that when referring to DLs the context is made very clear.

3.5.2 Analytical Capabilities and Project Requirements

Analytical capabilities are constantly improving, resulting in greater sensitivity and lower detection limits. This improvement in analytical capabilities is frequently the vehicle that drives regulatory, and hence, project requirements. Investigators often base their sensitivity requirements upon analytical method capabilities rather than upon project-specific objectives. The problem for the project team is to determine the levels of sensitivity needed to generate data adequate for decision making, to establish MQOs based on this evaluation, and to be sure that the indicator of sensitivity used to evaluate a particular method appropriately reflects the performance of the method in the particular matrix of interest. MQOs tie the required measurement quality to the project DQOs.

The project team should always consider the needed sensitivity of a measurement prior to requesting laboratory analyses. Once the needed sensitivity is determined, the project team can work with laboratory personnel to choose the appropriate analytical method.

3.5.3 Sensitivity Indicators

A large number of DQIs relate to sensitivity. Some commonly used examples include the instrument detection limit (IDL), method detection limit (MDL), practical quantitation limit

Example 3-12. Importance of Discussing Sensitivity Requirements

The allowable limit for formaldehyde in air (8-hour, time-weighted average) was recently lowered to 40 parts per billion (ppb). An analytical laboratory was asked if they could meet this limit. The laboratory's standard procedure involved analysis using GC/MS (without derivation) with a RL of 500 ppb. After investigating derivation procedures, the laboratory instead chose to use GC/MS with selected ion monitoring (SIM). Selected ion monitoring improves the sensitivity of the method because it allows more individual raw data points to be acquired, which are then averaged to provide the intensity displayed in the data system. Experimentation with SIM using the formaldehyde ion 29 (most intense) for quantification, and ion 30 for confirmation, showed that the laboratory could achieve these limits. The laboratory was able to successfully use 30 ppb as the lowest point on the calibration curve.

(PQL), and RL. These definitions can be based on a statistical estimate of the random error (instrument/method noise) in the measurement system at low measurement levels (i.e., MDL) or they can be the lowest measurement the instrument is capable of determining within defined parameters of precision and accuracy (i.e., PQL). These sensitivity indicators are only valid for the conditions (matrix, instrument, laboratory, method) under which they are derived. In order to communicate effectively with the laboratory, and therefore more precisely and meaningfully define MQOs, it is necessary to understand differences associated with each of these terms. A summary of these terms is presented in Table 3-3. A more detailed explanation of MDLs, PQLs, and RLs, plus a discussion of alternative. Approaches for defining method detection limits and other indicators follows the table. These definitions provide a basis for understanding laboratory definitions. Specific definitions need to be established for the project of interest and discussed with the laboratory performing the analyses.

3.5.4 Method Detection Limits (MDLs)

All analytical measurements are imprecise because they have inherent fluctuations that lead to uncertainty. Method detection limits facilitate the determination of whether a single observation represents a true signal (as opposed to noise). The definition of detection limits involves first establishing whether a signal has been identified, second, determining what the associated value for the signal is, and finally, determining whether the value is sufficiently far from zero that a decision can be made regarding the presence of the analyte.

Until recently, two approaches have generally been advocated to arrive at method detection levels. In the first approach, multiple aliquots of an analyte free matrix, or a true environmental matrix (i.e., a true, well-mixed sample of soil, sediment, waste water, etc.) spiked at a single concentration (close to the anticipated MDL) are analyzed. The resulting standard deviation, in concentration units, is multiplied by a statistically-derived factor (i.e., the MDL is therefore a multiple of the analytical standard deviation replicate analyses at a low-level

standard), to compute the MDL. The analytical standard deviation is assumed to be constant over a relatively short range of low concentrations.

In the second approach, a complete calibration at multiple concentrations is prepared. The resulting linear calibration model and confidence intervals on the regression line are used to define the MDL. More recently, sensitivity indicators have been proposed that are derived from the quality control samples routinely performed (e.g., blanks, matrix spikes), or result from looking at the bias and precision at multiple concentration levels, followed by spiking at a single level believed to be very close to the detection or quantification limit.

Table 3-3. Commonly Used Sensitivity Indicators

Sensitivity Indicator	Numerical Definition	Definition	Common Use
Instrument detection limit (IDL)	Defined by instrument manufacturer (e.g., 3 x background noise)	Lowest value at which *instrument* can distinguish from zero	Provides basis for determining an MDL
Method detection limit (MDL)	MDL = $t_{0.99}$ x s s = standard deviation of 7 replicates of a low level spike $t_{0.99}$ = 3.14	"minimum concentration of a substance that can be reported with 99% confidence that the analyte concentration is greater than zero"[1]	Determines the lowest quantifiable concentration for a given method
Practical quantitation limit (PQL)	PQL≈ 5 x or 10 x MDL (also can be determined or refined by including values near the MDL as the lowest standard on the instrument calibration curve)	"the lowest concentration of an analyte that can be reliably measured within specified limits of precision and accuracy during routine laboratory operating conditions"[2]	Provides numerical lower limit for critical data
Reporting limit (RL)	Laboratory defined (often; RL = PQL)	Lowest value reported by laboratory without a "J" flag	Laboratory basis for reporting data without any flags
"J" value	MDL < "J" value < RL	Estimated value; quantitated value below the laboratory RL and above the MDL	Alerting the data analyst that the value reported was below the laboratories' RLs.

[1] 40 CFR 136 Appendix B
[2] 50 FR 469906

A commonly used method for calculating the MDL follows the definition and procedures spelled out in 40 CFR 136, Appendix B. This method recommends analyzing seven replicates of

spikes close to the assumed MDL, calculating the standard deviation, and multiplying this by the t-statistic corresponding to a 99% probability of concluding the value is different from zero. The logic in picking a standard to run, is to obtain a standard deviation that is approximately one-third of the response of the standard under analysis (33% RSD, assuming a blank is zero response). This will result in an MDL calculation very close to the standard level analyzed (since the student's *t* value is very close to 3, depending upon the number of standards analyzed). If the laboratory could analyze a blank and obtain a standard deviation from that analysis, they would pick a standard that provides a response equal to three times the standard deviation of the blank. However, some methods are not amenable to blank analysis in this manner; often the signal of the blank is zeroed out either manually or automatically. The 40 CFR 136 definition is the most widely used; however, many others have been proposed and used because of the inherent confusion regarding the intended use and meaning of the term "detection limit." A brief example of the 40 CFR Part 136 approach is presented, followed by a description of a number of alternative approaches that have been proposed by various authors, associations and agencies.

The 40 CFR 136 method detection limit was based on a paper by Glaser et. al. (1981). The MDL was defined by Glaser as a concentration where an average 99% of the trials measuring the analyte concentration at the MDL must be significantly different from zero analyte concentration.

$$\text{MDL (in concentration units)} = t_{(n-1\ df,\ 1-\alpha = .99)}\ Sc$$

where Sc is the standard deviation from at least 7 aliquots (n) of a standard taken through the complete analytical procedure.

An MDL study using spiked samples should be performed at a concentration that is equal to or up to 5 times the level of the resulting calculated MDL. The value $t_{(n-1\ df,\ 1-\alpha = .99)}$ is chosen based on the number of aliquots analyzed, and the selection of alpha (Type I) error. Errors in calculating the MDL arise from improper selection of t and accepting calculated MDLs that vary by a factor of five [or ten depending upon the interpretation (see Rosecrance, 2000)] greater than the spiked concentration (e.g., spike at 100 and calculated MDL is 8).

Glaser et al. assumed normal distribution of the variance. They began the derivation of this indicator with the assumption that variability is a linear function of concentration. However, the resulting equation, as shown above, does not include variability in this light. They use the student's t-distribution to approximate the error distribution. The Glaser definition uses a low-level standard taken through the complete analytical procedure including preparation, clean-up, and instrumental analysis. The assumption is made that the distribution and variance of the low-level standard is equivalent to the variance of the blank.

This sensitivity indicator only accounts for Type I errors and is visually analogous to Currie's Lc. Glaser et al. provided guidelines (see above) in choosing the concentration level for this standard, but depending upon the calculated MDL, a series of experiments may need to be

Example 3-13. Method Detection Limit by 40 CFR 136

An environmental testing laboratory has purchased a new electrothermal atomizer atomic absorption instrument. The instrument is equipped with a new atomization technique marketed as more accurate (less interference), and potentially more sensitive. The instrument will be used primarily for water analysis. Following the procedure in 40 CFR 136 Appendix B, for method detection limit evaluation, the laboratory estimates the MDL to be 1.0 μg/L for a 10-μL sample injection. The results of the analysis of 10 aliquots of laboratory standards at 2 μg/L is shown below.

Measured Concentration	2.52	2.89	2.74	1.92	2.39	2.74	2.61	2.73	1.93	2.60
Aliquot	1	2	3	4	5	6	7	8	9	10

The variance and standard deviation calculated from the standard analysis are 0.113 and 0.336, respectively. Using the student's *t* value provided in the CFR (for 9 degrees of freedom) the MDL is calculated as:

$$MDL = t_{(9,0.99)} s = 2.821 \times 0.336 = 0.95 \quad .$$

Because this concentration is quite a bit lower than the 2.0-μg/L standard, the laboratory must re-perform the MDL study at a lower concentration. In the next iteration the laboratory picks a standard at 1.0 μg/L. The responses of the analysis of 7 aliquots of a 1.0-μg/L standard are provided in the following table.

Measured Concentration	0.70	1.36	1.25	1.03	1.25	0.69	0.87	Ave.	s
Aliquot	1	2	3	4	5	6	7	1.02	0.275

The student's *t* value provided in the CFR (for 6 degrees of freedom) is 3.14. Hence, the MDL is calculated as:

$$MDL = t_{(6,0.99)} s = 3.143 \times 0.275 = 0.86$$

With seven samples, the MDL calculation is 0.86 μg/L. This value is sufficiently close to the standard concentration of 1.0 μg/L to avoid another iteration in the MDL study.

performed to arrive at a satisfactory value. Because this definition for detection limit became the EPA defacto method, as codified in 40 CFR 136 Appendix B, it has widespread usage. Even though the method is well known, some authors have found it confusing as written, and note that essentially four different MDL values can result: an estimated, calculated, recalculated, and pooled value (Kimbrough and Wakakuwa, 1993). Kimbrough and Wakakuwa have results from MDL studies that were strongly biased and contained high numbers of both false positives and false negatives. Studies conducted by the Wisconsin Department of Natural Resources (WDNR, 1996 and Carden, 1998) reported that roughly half of 56 laboratories they looked at in 1993 incorrectly calculated the MDL, and that 26% of the results of MDL studies were incorrect in a survey conducted in 1998. These studies support the notion that there continues to be some confusion on how to properly implement the procedures specified in 40 CFR 136.

3.5.5 Alternative Sensitivity Indicators Related to Detection

A sampling of the various approaches to generating sensitivity indicators is presented below.

3.5.5.1 Minimum Detectable and Minimum Quantifiable Values.

The International Union for Pure and Applied Chemistry (IUPAC) has developed sensitivity indicators (Currie, 1995) that begins with defining the critical value (Lc) defined as the minimum significant value of an estimated net signal or concentration that can be applied as a discriminator against background noise.

$$Lc = z_{(1-\alpha)} \times \sigma_o$$

where $z_{(1-\alpha)}$ represents the (1-α)th percentage point or critical value of the standard normal variable, and σ_o is the standard deviation of the estimated quantity (net signal or concentration) under the null hypothesis (true value = 0). When σ_o is estimated, it must be replaced by s_o, based on ν (n-1, where n = number of replicates) degrees of freedom, $z_{(1-\alpha)}$ must be replaced by student's-t.

$$Lc = t_{(1-\alpha),\,\nu} \times s_o$$

The critical value is based on a statistical test for deciding between the hypothesis that the sample contains no analyte versus the alternative, the sample contains an analyte (true concentration is greater than zero). Only Type I errors (the probability of falsely concluding an analyte is present, also known as a false positive) are considered.

To address Type II errors (the probability of falsely concluding an analyte is not present, also known as a false negative), the minimum detectable value (Ld) is defined as that value (Ld) for which the false negative error is β, given Lc. It is the true net signal (or concentration) for which the probability that the estimated value does not exceed Lc is β.

$$Ld = Lc + z_{(1-\beta)} \times \sigma_D$$

Assuming constant variance, and letting α equal β, then

$$Ld = 2\ Lc$$

Again, when σ_d is estimated, it must be replaced by s_o, based on ν degrees of freedom, $z_{(1-\beta)}$ must be replaced by the non-centrality parameter of the non-central-t distribution.

Following Currie's definition, a minimum detectable value requires incorporating the criteria for accepting both Type I and Type II (false negative) error. Therefore alpha (for Type I) and beta (for Type II) must be chosen *a priori*. Often, alpha is set equal to beta, and the minimum detectable value is approximately equal to (for constant variance):

$$Ld \approx 2t_{1-\alpha,\nu}\ \sigma_o$$

If only an estimate S_o is available (for σ_o), the minimum detectable value is uncertain by the ratio (σ/s). Currie provides correction factors for 2t given that the number of samples (and degrees of freedom) are generally less than 25.

One must realize that if a RL is set at the Ld, the Type II error acceptance rate defaults to 50%, because by censoring at the Ld, 50% of samples that are truly at Ld would be reported as nondetects.

Having defined the critical value and minimum detectable value, Currie then goes on to define the minimum quantifiable value (also referred to as the quantification limit) (L_Q) as a performance characteristic that marks the ability of a chemical measurement process to adequately "quantify" an analyte. The ability to quantify is generally expressed in terms of the signal or analyte (true) value that will produce estimates having a specified RSD, commonly 10%.

The three IUPAC definitions assume that the concentrations under study follow normal distribution. Blanks are used to obtain the estimate of standard deviation and the reference includes a good discussion concerning the different types of appropriate blanks, a very crucial component of these indicators.

Figure 3-7 shows the relationship of the three indicators defined by IUPAC. The non-constant variance (different shapes of the distributions) is purposely included. In this Figure, alpha is less than beta, and the quantification limit is set at 10% RSD. The development of detection limits in the environmental arena has generally followed the logic outlined in this derivation. This definition does not prescribe the minimum number of data points required to develop the sensitivity indicator.

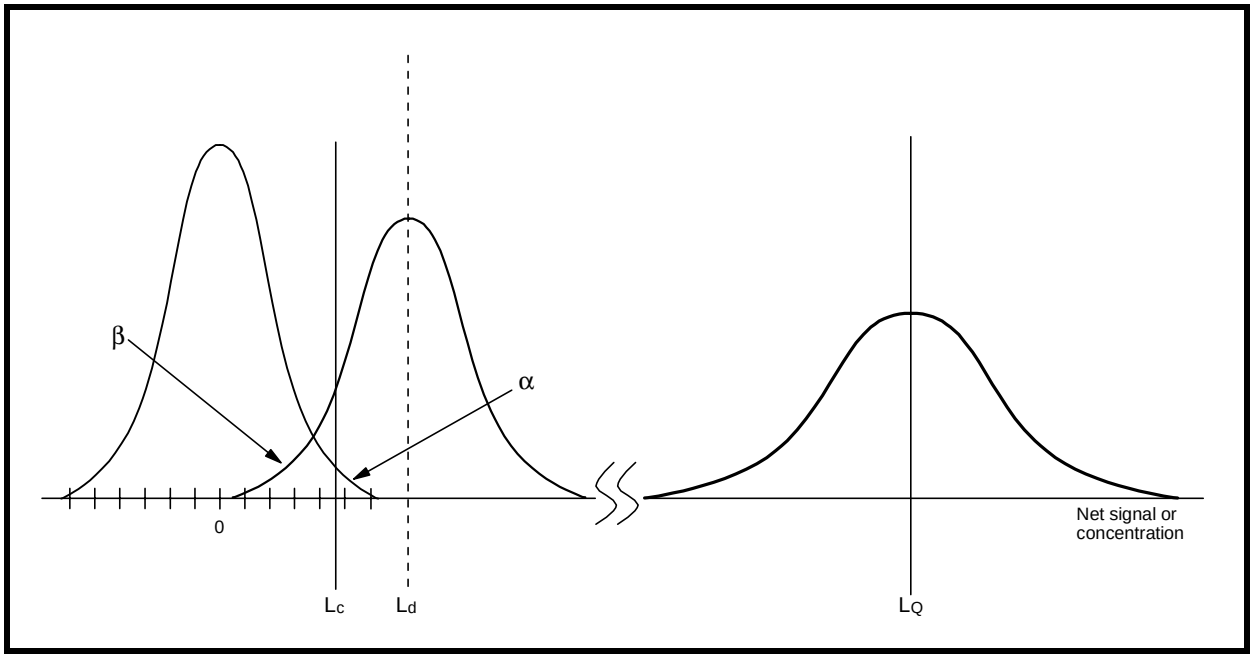


Figure 3-7. The IPAC Definitions for Critical Value (Lc), Minimum Detectable Value (Ld), and Minimum Quantifiable Value (L_Q)

3.5.5.2 Limit of Detection, Reliable Detection Level, and Limit of Quantification

L. H. Keith (1991) developed three separate sensitivity indicators that account for background signal from a blank: the limit of detection, reliable detection level and limit of quantitation. Each of these indicators are summarized below.

Limit of detection (LOD) is defined by Keith as the lowest level that can be determined to be statistically different from a blank at a specified level of confidence. This corresponds approximately to Currie's "critical level" (the standard deviation is simply multiplied by three, instead of multiplying by $t_{(1-\alpha),\ v}$).

$$LOD = b + 3\sigma$$

where b is the average signal from the blank and σ is the standard deviation of the blank, which is similar to the approach taken by IUPAC/Currie.

Reliable detection level (RDL) is defined by Keith as the concentration level at which a detection decision is extremely likely. The RDL is generally set at a higher level than the MDL or LOD. As is evident, this would be close to the IUPAC/Currie minimum detectable value since it corresponds to twice the LOD after the blank, or background level is accounted for.

$$RDL = b + 6\sigma$$

Limit of quantitation (LOQ) is defined by Keith as the level above which quantitated results may be obtained with a higher, specified degree of confidence.

$LOQ = b + 10\sigma$

Keith indicates that the LOQ corresponds to an uncertainty of ±30% in the measured value at the 99% confidence level. The LOQ can be varied with the desired confidence level. 10σ is chosen to provide a signal to noise ratio of 10:1 or RSD of 10%. This is analogous to the IUPAC/Currie minimum quantifiable value.

Keith chose to use constants, instead of choosing a "t" value based on the degrees of freedom and alpha and beta levels. Keith indicates that the constants he has chosen provide false positive and false negative probabilities of approximately 0.1% at the RDL level.

It should be noted that if the RDL is used as a RL (wherein values below the RDL are censored and reported as less than the detection limit), the false negatives probability becomes 50%. This is because a sample that is truly at the RDL level will fall (be calculated) below the RDL 50% of the time, due solely to chance. These indicators provide a very convenient form of calculation, since constants are used. However, there is some concern that the simplifications and substitutions for convenience may over simplify what is required to arrive at reliable indicators. The issues of normal distribution, constant variance, and estimation of variance using different student t-distributions are all hidden by Keith's approach. There is growing evidence that these issues cannot be assumed in many cases.

The major disadvantage of single concentration designs is that an assumption must be made; that variability at the spiked concentration is identical to variability at the true MDL. Another criticism of these traditional data indicators is the lack of accuracy or precision information near the detection limit. They are relatively easy to perform and calculate and have large recognition in the environmental laboratory community.

3.5.5.3 Decision Limit and Detection Limit

Hubaux and Vos (1970) introduced the first algorithm for obtaining estimates of what they called the decision limit and detection limit based on the on the upper and lower prediction intervals derived from a least squares fitting of a calibration data set. With *a priori* choice of the prediction intervals, they derive threshold values similar to a decision and detection limit.

Decision limit (Up) (Figure 3-8) is defined by Hubaux and Vos mathematically as:

Yp (in instrument response units) = $a + t_{1-p} s \sqrt{1 + \frac{1}{n} + \frac{\overline{X}^2}{Q^2}}$

and Up (in concentration units) is equal to: (Yp-a)/b. This is where the horizontal line from Yp crosses calibration line.

a = estimated intercept on the Y axis of the calibration line.
b = estimated slope of the calibration line.
t_{1-p} = (1-p)th percentile of the t distribution with n–2 degrees of freedom.
s = root near square error for fitted calibration line.
n = number of calibration samples used to estimate the calibration line.
$\overline{X}$ = average of the concentrations of calibration samples.
$Q^2 = \sum \left(X - \overline{X} \right)^2$ for calibration samples.

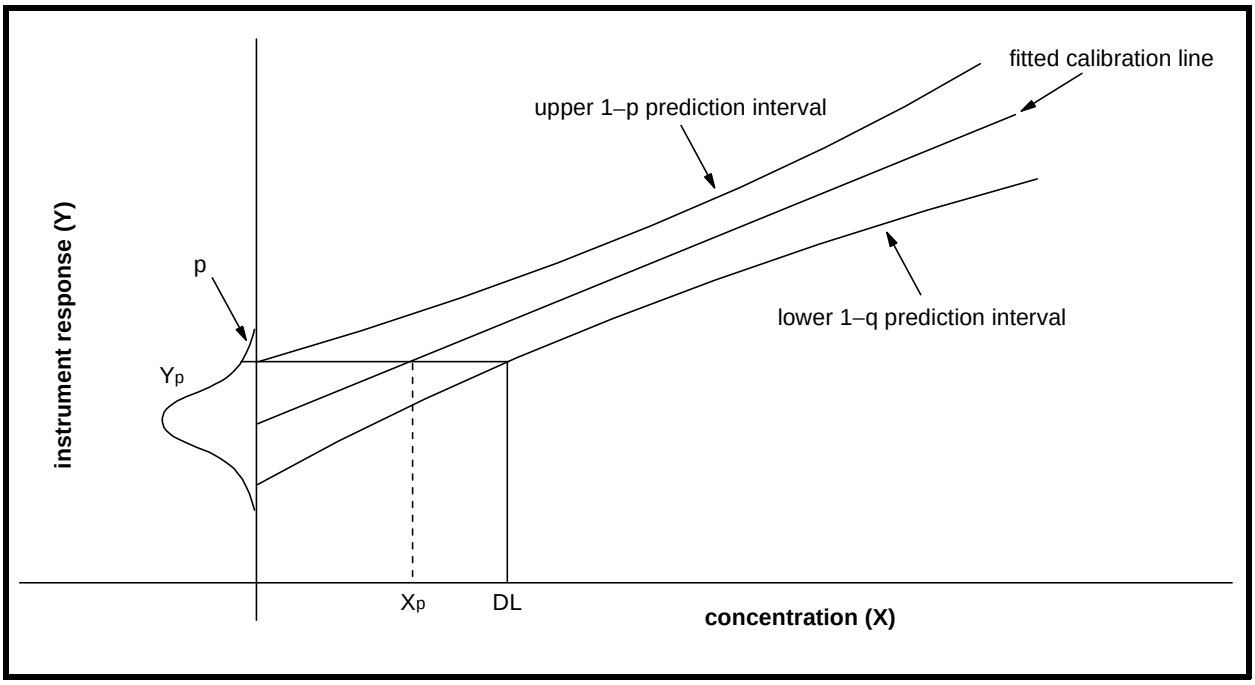


Figure 3-8. Decision Limit (Up) and Detection Limit (DL) Derived from Calibration Prediction Intervals

Detection limit (DL) (see Figure 3-8) is obtained by matching the horizontal line from Yp to the lower qth percentile of the prediction formula.

The decision limit is based on Type I error only and in this way is analogous to Currie's (Lc). However, Hubaux and Vos utilize the prediction intervals from the regression line to estimate the uncertainty. The value of the Detection Limit is influenced by not only the inherent variability of the method, but also the number of samples used for the calibration and the calibration concentrations. However, constant variability through the calibration range is assumed.

Hubaux and Vos's algorithms for determining sensitivity indicators could, in theory, utilize the initial calibration curve required for most analytical procedures instead of a separate detection limit study. More attention would need to be given to the number of calibration levels and replicates at each level so that the "lack of fit" and "pure error" statistics would be available to evaluate the calibration curve. The numerical calculation is clearly more complex than the indicators defined above; however, this could be automated. Unlike Glaser et al., the standards used in this derivation are not normally taken through the complete analytical procedure; however, these could be extended to include sample processing and using true matrices, such as the method of standard addition.

3.5.5.4 Detection Limits with Specified Assurance Probabilities

Clayton (1987) refined the procedures developed by Hubaux and Vos. Like Hubaux and Vos, Clayton uses calibration regression to arrive at detection limits. However, Clayton employs a non-central t-distribution because the assumption of normal error distribution is only valid for the null hypothesis (is the true value equal to zero), not the alternative hypothesis (is the true value greater than zero) (Gibbons, 1994).

Detection limit was formulated by Clayton in terms of the concentration at which the probability of choosing Ct > 0 over Ct = 0 is 1-q.

Yp and Up (like Hubaux and Vos) are referred to as threshold values.

$$\text{D.L.} = \Delta \; s \frac{\sqrt{1 + \frac{1}{n} + \frac{\overline{X}^2}{Q^2}}}{b}$$

Δ is a non-centrality parameter of the distribution corresponding to p,q and degrees of freedom in the t-distribution. (n–2, where n is the # of samples in the calibration line). Table of Δ values was provided by Clayton in the paper. As n increases, the difference between central t and non-central t becomes negligible. All other symbols are identical to those above used by Hubaux and Vos.

The detection limit, as defined here, permits the analyst to specify both Type I and Type II error rates. The non-central t-distribution can also be employed in the IUPAC/Currie indicator for Type II error.

3.5.5.5 Detection Limits Based on QC Data

Osborne and Rocke (2000) propose a two-component model to calculate measurement error using the quality control data routinely available (blanks, matrix spikes).

(1) $$Su = \sqrt{\left[S_o^{\,2} + u \times e^{RSD^2}\left(e^{RSD^2} - 1\right)\right]}$$

where Su = standard deviation, So = standard deviation of blank, u = average concentration, RSD is relative standard deviation.

By combining this with the 40 CFR 136 definition of the MDL (or another definition if desired).

(2) $MDL = t \times s$

The detection limit based on QC data is equal to the concentration that satisfies both equations. To solve these equations, an initial estimate of the detection limit is used for u. The RSD is obtained from replicates at high concentration or from batch QC samples (MS/MSD). So is estimated from method blank data. Equation (1) is then solved for Su, the result is plugged into equation (2) to compute the MDL. This result is then plugged back into equation (1) until both equations converge.

The Osborne and Rocke model approximates a constant standard deviation model for very low concentrations and approximates a constant coefficient of variation (relative standard deviation) model for high concentrations. Because of this, Osborne and Rocke claim that their approach provides a more robust sensitivity indicator. The Osborne and Rocke approach avoids costly bench time required to conduct a separate detection limit study. Their approach provides potential for more reproducible detection limit estimates with tighter confidence limits by using more than 7 measurements for the origin of the uncertainty. Because quality control samples are employed, it includes the variance associated with the complete analytical method, and can readily be updated over time. The model can also be applied to non-linear calibration, censored data, and the method of standard additions. However, like the other traditional indicators discussed above, no bias or precision information is included. This indicator would benefit from additional work that utilizes real laboratory data to compare it with other estimators.

3.5.5.6 Instrument Quality Control Level and Method Quality Control Level

In this approach, the authors (Kimbrough and Wakakuwa, 1994) advocate reporting concentrations on acceptable precision and bias, which they assume are a function of concentration. They show that precision and bias can change dramatically over a very small concentration range indicating that estimators based on an arbitrarily chosen concentration can be inappropriate.

Instrument quality control level (IQCL) is defined by Kimbrough and Wakakuwa as the smallest concentration of an interference-free standard analyzed on a particular instrument that is within the quality control and assurance parameters of the laboratory and data user. For example, an IQCL can be stated as the concentration where 50 ±5% bias and RSD can be achieved. To

determine this concentration, replicate standards are analyzed at multiple levels and bias and precision are evaluated. The lowest concentration that meets (or exceeds) the precision and bias criteria is selected.

Method quality control level (MQCL) is defined by Kimbrough and Wakakuwa as the smallest concentration of an analyte in a given matrix by a given method and instrument that is within the DQOs of the end user for the data and the laboratories' quality control parameters. The MQCL is estimated by performing the following steps sequentially (each provides an estimate of the MQCL):

1. Multiply the IQCL by the correction factor that accounts for analytical procedures, such as extraction, concentration, and digestion.

2. Spike the matrix of interest at 3 levels or more, including at the estimated MQCL, below, and above this level. The lowest spike level that provides the highest acceptable bias and variance (as determined by the DQOs) is the second estimate.

3. Using the same matrix, prepare a standard at the second MQCL estimate, analyze this sample 7 times. If the resulting bias and precision are acceptable for this concentration level, it should be chosen as such. If it is not, the procedure should be repeated until an acceptable concentration is determined.

The MQCL should be confirmed by analysis of a different standard at the estimated MQCL. Data should be reported as follows:

- If the results of an unknown are greater than the MQCL, report the number, along with the bias and precision of the MQCL.

- If the unknown is less than the MQCL, report as "not detected in quantities greater than MQCL."

- If the unknown is greater than the MDL, but less than the MQCL, state this in prose, without reporting a value.

These indicators provide a very specific laboratory, instrument, and perhaps analyst dependant estimate. They require considerably more preparation and analysis time than other sensitivity indicators including the MDL as defined in 40 CFR136. The censoring advocated by the authors would potentially (depending upon the DQOs and how the data is applied) result in lost information for later data interpretation. The advantages include both a precision and bias that meets the project objectives, and the confirmation of these values with an autonomous standard.

3.5.5.7 Interlaboratory Detection Estimate and Interlaboratory Quantitation Estimate

The approach taken by the American Society for Testing and Materials (ASTM, 2001a and b) focuses on laboratory performance that is representative of routine measurements attainable by most quality laboratories. Thus, the sensitivity indicators encompass variance from multiple laboratories, instruments, and even matrices to some extent. Data (multiple concentrations) from at least six different laboratories are used in the indicator derivation.

Interlaboratory detection estimate (IDE) D6091 as defined by ASTM (2001a) "is computed to be the lowest concentration at which there is a 90% confidence that a single measurement from a laboratory selected from the population of qualified laboratories represented in an interlaboratory study will have a true detection probability of at least 95% and a true nondetection probability of at least 99% (when measuring a blank sample)."

Ideally, the study entails multiple laboratories analyzing spiked and blank samples completely blind, with no knowledge of the concentrations or special nature. Multiple levels (at least 5) are analyzed, with blanks and the concentrations such that at least one spike is approximately twice the anticipated IDE and one below. The resulting set of data is evaluated for an appropriate model (constant, straight line, curved) and YC (the measurement critical for detection) is chosen base on alpha and beta (for Type I & II errors, respectively). The limit of detection is a recursive function that is solved iteratively. Graphically, the result is shown in Figure 3-9.

The ASTM approach uses Currie's logic for picking alpha and beta, but goes farther by realizing that the relationship between measured values and true concentrations must be modeled. The variation introduced by different laboratories, analysts, and other factors are all included in the calculation.

Interlaboratory quantitation estimate (IQE) D6512 as defined by ASTM (2001b) "is computed to be the lowest concentration for which a single measurement from a laboratory selected from a population of qualified laboratories represented in an interlaboratory study will have an estimated Z% relative standard deviation (Z% RSD, based on an interlaboratory standard deviation), where Z is typically an integer multiple of 10, such as 10, 20, or 30, but Z can be less than 10."

The same data obtained for deriving the IDE is used to arrive at the IQE. Using the interlaboratory standard deviation model to estimate the true interlaboratory standard deviation, and the mean recovery to scale the standard deviation, the IQE is found by solving the following equation:

$$T = (100/Z)*G(T)$$

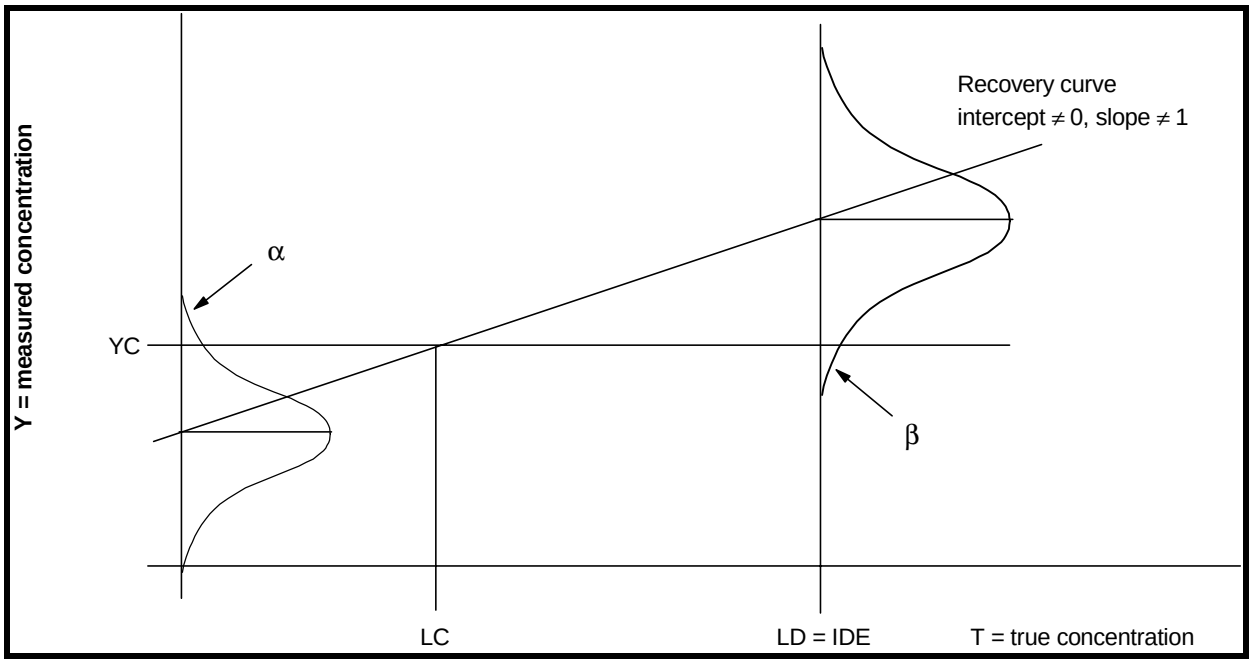


Figure 3-9. Critical Level (Lc) and Interlaboratory Detection Estimate (IDE) from ASTM

G(T) is a function that estimates the interlaboratory standard deviation (in concentration units) of the true concentration (T). Z is the desired precision (increasing 10, 20, 30).

IQE is analogous to limit of quantitation after Currie, with an RSD of $\leq$10% considered a quantitative approach.

The indicators developed by ASTM provide algorithms to generate detection and quantification at which "any qualified laboratory" should be able to reliably detect with few false positives (with X% confidence that the analyte is detected at DL and blanks measured with Y% of false positives) and quantitate at a specified precision (RSD). The results of using these indicators and sending out PE samples has shown that:

- most labs have a low rate of false positives (blind blank samples),
- most labs can reliably detect at the IDE, and
- most labs can report measurement values with known, limited RSD at (and above) the IQE.

Calculating detection limits based on the ASTM interlaboratory approach assures that the indicators are attainable by most laboratories. It includes multiple sources of variance and is

therefore more realistic. The resulting indicators provide estimations of bias, precision (RSD), and the associated confidence. Data from at least six different laboratories is required, which is difficult in the present environmental laboratory structure. The approach outlined here would prove useful in examining the ability of laboratories to meet regulatory limits (e.g., discharge levels) when the risk based limit (e.g., MCL) is near or below the perceived detection limit.

3.5.5.8 Long-Term Method Detection Level and Laboratory Reporting Level

The United States Geological Survey has developed two sensitivity indicators that take a perspective similar to ASTM. Long-term, multiple sources of variability are included to yield indicators that are at a higher concentration than would generally be obtained at a single laboratory.

Long-term method detection level (LT-MDL) as defined by USGS is calculated as follows:

$$\text{LT-MDL} = t_{(n-1\ df,\ 1-\alpha = .99)} \times Sc$$

This is a modification of EPA MDL as defined in 40 CFR 136, Appendix B. The equation is identical, however a larger number of replicate spike samples are used ($n \geq 24$) and these are collected over an extended period of time (6 to 12 months). All aspects of routine analysis (multiple instruments, operators, calibrations, and preparation) are included. The assumptions are identical to Glaser and Keith, near-normal (t-distribution) distribution, and constant variance from LT-MDL spike concentration to zero concentration. The resulting value is a "best-case" detection limit, because they use reagent water to spike.

The USGS National Water Quality Laboratory collects uncensored blind blank data for many methods. This data can be used to compute the LR-MDL directly, and represents a simpler alternative than the use of spikes.

If a fixed blank bias is known, it must be accounted for using the following equation:

$$\text{LT-MDL} = A + t_{(n-1\ df,\ 1-\alpha = .99)} \times Sc$$, where A is the median (or mean) blank concentration.

As of 2001, Sc has been replaced by Fσ (LT-MDL = T x Fσ), where Fσ is pseudosigma. Pseudosigma is defined as the interquartile range of the data divided by 1.349 (Hoaglin et al., 1983).

Laboratory reporting level (LRL) is defined by USGS as follows:

LRL = Z x LT-MDL, Z = 2/ recovery of LT-MDL spikes.

Z is equal to a factor of 2 or more to give a ≤1% false negative rate at LRL. Ideally, the lowest calibration standard is equal to the LRL. Reporting of data is dependant upon methods used in the analysis. (See http://wwwnwql.cr.usgs.gov/Public/ltmdl/ltmdlsplash.html)

The USGS plans to use these estimators for all National Water Quality Laboratory operations. They intend to use uncensored blind-blank sample to obtain LT-MDL and allow for blank off-set corrections.

A large component of these new indicators is a change in philosophy in setting the censoring level. No censoring will occur down to the LT-MDL level, though results between the LRL (or lowest standard, whichever is greater) will be qualified. In addition, information-rich techniques that can provide additional data to be used in identification (e.g., spectral techniques) will not result in censoring any data that has been "called" by the analyst. In fact, any censoring will depend upon the project data-quality objectives.

The numerical value of the LT-MDL is generally greater than the EPA MDL because more and larger sources of variability are measured as a result of using multiple instruments and calibrations, and because this information is obtained over a greater time period. This approach is similar to ASTM in that the variance is associated with multiple laboratories (and instruments, calibrations, analysts). However, the calculations as presented by USGS are much more straightforward than ASTM. The LRL also assumes near normal distribution, in contrast to the approach of Currie and Clayton et al.

By using blank analyses routinely performed, the necessity of extended bench time for detection limit studies is minimized. However, they have identified analytes that show non-constant variance between the blank and spike levels, resulting in fairly large differences in the resulting LT-MDL calculation.

3.5.5.9 Alternative Minimum Level

Gibbons et al. (1997) recently introduced another sensitivity indicator that has drawn considerable interest (see Gibbons et al., 1977; Kahn et al., 1998; Gibbons et al., 1998; Rigo, 1999; Gibbons et al., 1999 and Kahn et al., 1999).

The alternative minimum level (AML):

AML = X_q + (t/b_{1w})square root[$v(y_Q)$] (see reference for derivation)

The AML is an extension of Currie's Lq. It accounts for the observed non-constant variance to provide a solution that will typically yield a RSD of 10%. Gibbons et al. apply a model to describe the variability with concentration obtained from a multiple level calibration. There has been considerable discussion about the validity of this model (see references below).

There is considerable computation required to arrive at the AML and alternative approximations for both quantification and detection indicators have been presented by Zorn et al. (1999).

3.5.6 Instrument Detection Limits

Instrument detection limits (IDLs) are an indicator of the concentration of a constituent that an instrument can distinguish from zero. IDLs are typically based on an analysis of blank samples to determine the background noise of the instrument, and may be defined as three times the noise level. They do not take into account method-specific variation.

Generally, IDLs are provided for inorganic instrumentation (e.g., ICP) but are not commonly used terminology for organic instrumentation such as a GC or GC/MS. They are usually defined by the instrument manufacturer as the lowest level a particular instrument is capable of measuring. The IDL is not rigidly defined in terms of matrix, method, laboratory, or analyst variability, and is not usually associated with a statistical level of confidence. As such, IDLs are usually lower than MDLs and rarely serve a purpose in terms of project objectives. They are often used by the laboratory as a means for deciding upon the purchase of instrumentation or as a starting point in deriving MDLs.

3.5.7 Practical Quantitation Limits

The requirements for quantification are more stringent than for detection. A quantification limit defines the concentration at which a method provides a specified precision. To an analytical chemist, this is usually considered 10% RSD. Thus, limit of quantification (LOQ) is generally defined at 5 to 10 times the standard deviation of the noise (or blank) signal. Because the MDL is approximately 3 times the noise signal, if a factor of 5 or 10 times the MDL is used to calculate the PQL, this will exceed the requirement of 10% RSD.

In practice, PQLs are usually defined at the lowest concentration in the calibration curve. When this approach is used, the calibration statistics (e.g., correlation coefficient, confidence intervals, RSD of response factor) ensure that the PQL represents the same precision and accuracy as other data reported for the analyte. Using the lowest standard of the calibration is routine for organic analysis.

Practical quantitation limits are important DQIs for sensitivity, because they represent the lowest value that can be quantified with an acceptable degree of confidence. An MQO may be set quantitatively for the PQL, such that an inability to achieve that sensitivity would cast doubt on the attainment of the DQOs.

Example 3-14. Practical Quantitation Level for Arsenic

In addition to laboratory or project PQLs, regulatory agencies, such as the U.S. EPA, have derived PQLs through interlaboratory studies. This practice leads to a range of different PQLs being set by each laboratory. The EPA recently published a new PQL for arsenic at 3 μg/L in water, with an acceptance limit of ±25-30%.

3.5.8 Reporting Limits

Reporting limits are project or laboratory specific. Laboratories often develop RLs based on the needs of the client, or may simply default to the laboratory PQL. Laboratory RLs should change with the projects' needs and not be set simply by the laboratories' standard procedures. The relationships of MDL, PQL, LOD, and LOQ to the instrument signal from a blank are shown in Figure 3-10. As the signal level increases (as a multiple of the standard deviation from a blank), the confidence in detection and quantification improves. At the MDL (as defined in 40 CFR 136) there is a 99% confidence that the signal measured represents detection. Between the MDL and the PQL, there is less confidence in the estimate of the analyte concentration.

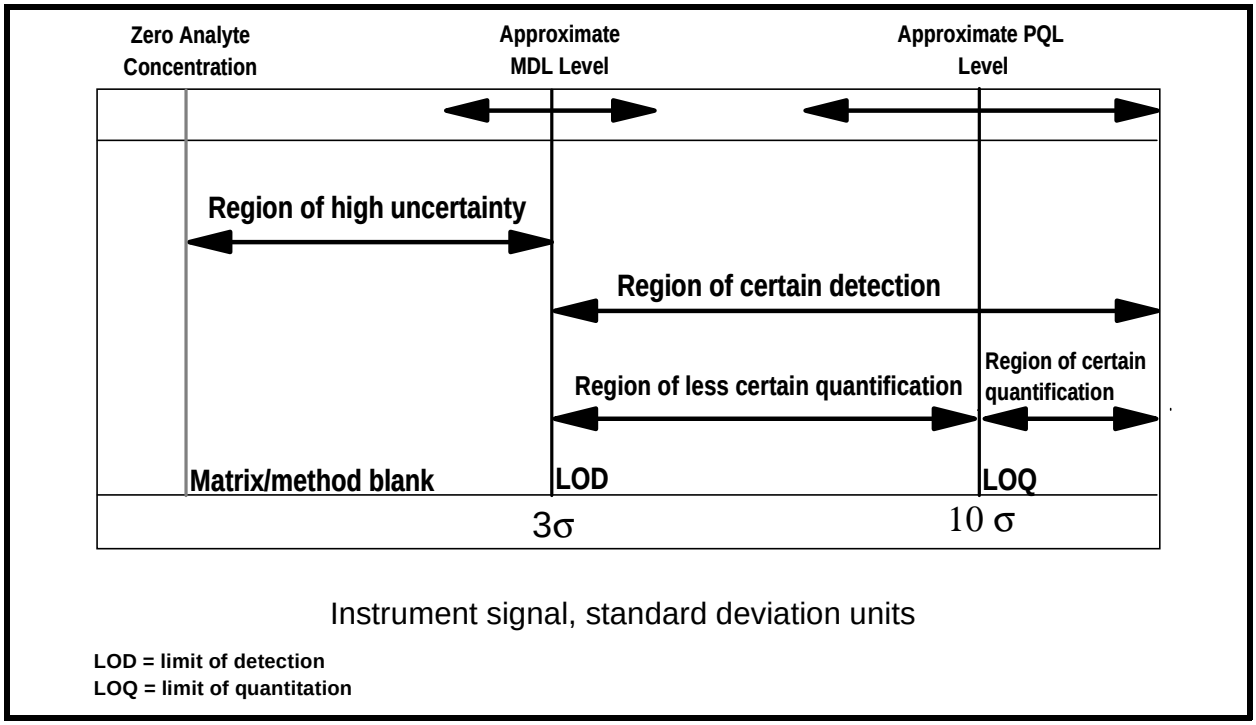


Figure 3-10. Instrument Signal Levels

3.5.9 Selection of the Appropriate Sensitivity Indicator

Selection of the appropriate sensitivity indicator is important for evaluation of project objectives. Prior to project initiation, the project DQO team should establish levels of detection that will be needed, determine if the method sensitivity will meet the project needs, and determine if the PQL or the MDL will be used to evaluate the primary objectives. It is usually undesirable to determine critical project objectives using only estimated values.

MDLs calculated using the 40 CFR 136 method should be used for determining the lowest level that can be measured and should be qualified appropriately. The PQL, confirmed as the lowest point on the method calibration curve, may be used as the lowest value for evaluation of project objectives. The RL will be defined based upon the sensitivity requirements noted for the project. During development of DQOs, the project team should determine the concentration for the required sensitivity for each analyte, and a precision that must be associated with that concentration. Figure 3-10 provides a basis for considering the required sensitivity.

3.5.10 Confidence and Reported Data

As measurements are made approaching the PQL, MDL and below, the confidence associated with those data diminishes. Therefore, in determining whether more or less sensitivity is needed, the following types of questions should be considered.

Example 3-15. Water Versus Soil MDL

Why is there so much difference in the MDL (and PQL) between soils, sludges, and waters? Partly because of the different units, μg/L versus μg/kg. But, more importantly, it is a function of sample size, interferences from the matrix, and laboratory experience. There are techniques for lowering the detection limits, but they take time to develop and are more costly. If the quoted MDL/PQL is not satisfactory, alternate laboratories should be considered.

- How near the action level is the detection limit?
- Will one or more levels of dilution be necessary for the different analytes?
- Are these data being collected for general information or scoping, or for a more complex analysis like baseline risk assessment or site closure

3.5.11 Sensitivity and Measurement Confidence in Terms of Precision and Accuracy

There is always an instrument range where values reported are only rough estimates but are known to be above the noise level of the instrument. The analyst can determine that the compound is above background but the reported concentration will not have a known precision and accuracy associated with the reported value. These values are often noted with a "J" qualifier. This "J" value may be appropriate if the answer needed is whether or not a compound is present and to determine its approximate concentration. Caution should be applied when using "J"-flagged values for regulatory decisions or other uses, such as judging technology performance. It is strongly recommended that all instrument readings be reported and appropriately flagged, rather than censoring the data set at the RL or other level. In this way, a statistical and scientific judgment can be made as to how to use these results, and no information is lost.

Sensitivity indicators are defined to ensure positive identification of a compound. As noted by the equation used for MDLs, there is very little chance that if a laboratory reports a compound as being detected, that the reported compound is not present in the matrix.

Previously presented definitions for each of the indicators show that while positive compound identification can be assured, precision and accuracy associated with the reported value can only be assured if the lowest desired value is included in the calibration curve, as recommended for the PQL.

IDL ➜ MDL ➜ PQL

➜ increasing precision and accuracy ➜

The accuracy and precision associated with the reported value will be different depending on the laboratory definition for MDL and PQL and the data qualifier used for the reported data. Table 3-4 shows common laboratory qualifiers and their relation to the sensitivity indicator.

Table 3-4. Common Laboratory Qualifiers

Data Qualifiers	Condition	Relation
U	Non-detect	MDL > Result
J	Estimate	MDL < Result < PQL
None	Acceptable precision	Result > PQL

3.5.12 Project Needs Versus Analytical Potential

Regulatory or project requirements may be based upon instrument sensitivity potential. A prime example of this is the determination of complete cleanup of a spill of Resource Conservation and Recovery Act (RCRA) listed hazardous waste in soil. Choice of method (e.g., GC versus GC/MS) should be based on matrix and budget constraints, and may be a matter of allocating funding. For example, although an analytical method with more sensitivity may be required for soil samples, water samples may achieve the required sensitivity from a different, less sensitive, method. Less sensitive methods may be less costly. The primary concerns are to ensure sensitivity is aimed towards project requirements (MQO selection process), determine if method sensitivity achieves these requirements, and determine if the method being used for analysis is within budget constraints.

Consider, for example, organic analyses. Several factors are important to consider in selection of an analytical method, including the sensitivity of the method given, the sample matrix, whether the method is approved by the EPA, the full range of analytes required, and the cost per analysis. In many instances, lower detection limits for organics are achievable using GC-electron capture detector rather than using GC/MS. Other factors that can influence method performance is the sample matrix and matrix interferences. If interferences for the compound(s) of choice exist, than performing a GC analysis (without MS confirmation) may not be suitable. While GC/MS is often less sensitive (e.g., lower limit of detection 10 times higher than a GC method), it may be the only method of choice for a dirty matrix, because it is less prone to bias from interference. The effect of matrix upon sensitivity can be represented as follows:

drinking water > groundwater > soil or sediment >> hazardous waste

➜ increasing MDL ➜

When making decisions regarding method selection, the intended use of the data should also be considered. Rather than simply selecting the most sensitive method available, the project requirements should be balanced with analytical issues of cost and confidence. One should also consider that detection limits for a given method will vary from one matrix to the next. If information about the project, such as potential interferences (e.g., petroleum contamination, high level of non-analyte metals), is known, an appropriate analytical method should be chosen.

Measurement quality objectives for sensitivity are developed based on project objectives (e.g., what lower limits of detection are needed to evaluate a technology). Matrix, budget, and laboratory choice will also play an important role in determining what method to use in relation to needed sensitivity or level of detection and quantification.

3.5.13 Practical Quantitation Limits/Censoring

Practical quantitation limits are used to indicate the concentration of an analyte that provides a specified degree of precision and accuracy. Again, accuracy is only associated with the PQL level if it has been derived from the lowest standard used in calibration.

Values below PQLs, but greater than MDLs, are generally flagged with a "J," indicating they are estimated values; however, there is adequate certainty (e.g., 99%) that they are indeed valid detects. Reporting the estimated values, rather than censoring these results, is strongly recommended. When used to estimate averages or associated statistics, the estimated value is often deemed more appropriate than using substitution methods such as half the PQL. This is particularly true when estimating variances, because censoring data at the detection limit artificially lowers the variability. Values below detection limits are generally reported by the laboratory as below a specified value and flagged with a "U" (it is recommended that the laboratory be required to specify the MDLs for this purpose). Computing statistical estimators with data below a specified MDL can result in a skewed data set. [See Chapter 4.7 of *Data Quality Assessment (EPA QA/G-9)*(U.S. EPA, 1996)].

3.5.14 Communication

An iterative process may be required when project participants resolve sensitivity issues. The project DQOs may require that a regulatory limit be achieved, a percent decrease be measurable with a desired degree of confidence, or an estimate of some lower concentration be measurable with a designated degree of precision and accuracy. For projects where the DQOs include meeting a regulatory threshold, establishing an MQO for minimum sensitivity is relatively easy. For other cases, it is more likely that establishment of the sensitivity MQO and associated DQIs will require focused discussion and input from the statistician, project chemist, QA Officer, field team, and analytical laboratory.

How Sensitivity Is Communicated to the Laboratory

Early communication with the laboratory can determine if the required sensitivity is met by the laboratory's standard operations for the matrix of concern. Matrix, potential interferences, and required turnaround times all play a role when evaluating whether project sensitivity requirements can be achieved.

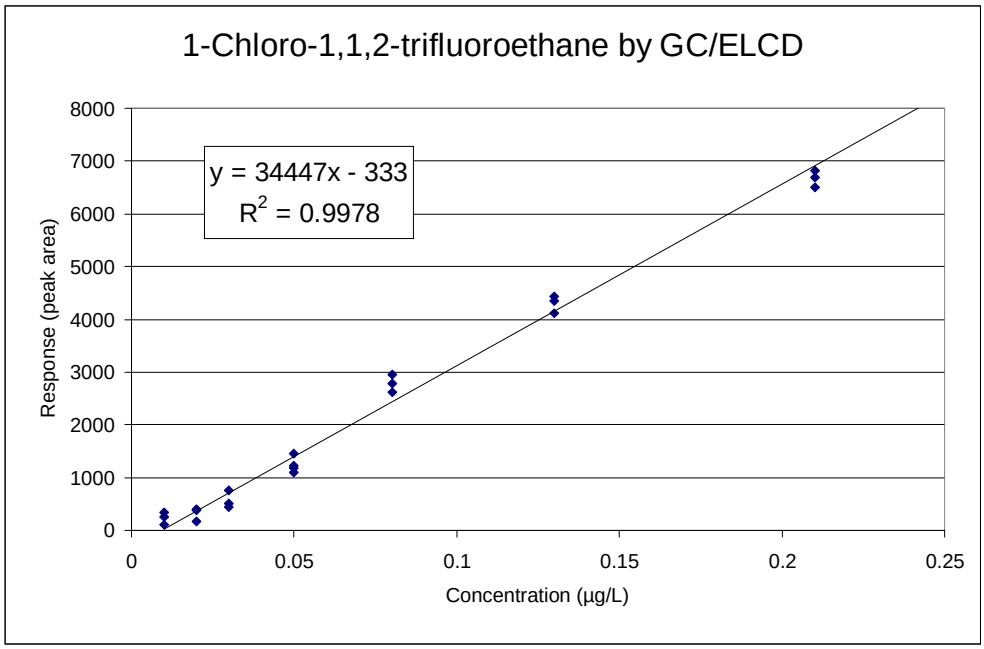
When the required sensitivity for meeting project DQOs is near the laboratory's capability, extended communication between project personnel and the laboratory is required. Generally, obtaining increased sensitivity with respect to detection issues causes a decrease in precision. Several factors can be discussed with the laboratory to control sensitivity, including method selection, adding initial calibration standard(s) at and below the required sensitivity level, and performing sample concentration.

Laboratory personnel will discuss sensitivity from an instrument calibration perspective. Their role is to determine what methods can be used to obtain the sensitivity needed. Additional issues that must be considered include the precision and accuracy requirements at various concentrations of interest, including the RL. Instrumentation capabilities will vary. It is important to remember that a different laboratory may offer more choices in methods that can be used to obtain better sensitivity. Not all laboratories are equal in this aspect of desired capabilities. The project requirements should determine the method used for analysis. For example, GC/MS, often the laboratory method of choice because of the laboratory's need to fit many different project specifications, offers more positive identification of compounds but can be less sensitive than GC methods using other detectors. The laboratory may try to make the method fit the project requirements rather than choosing the method based upon project requirements. A choice of GC/MS should be based upon project considerations. Because GC/MS is commonplace and has gained a reputation as "the better method," more laboratories perform organic analysis by GC/MS than by GC. For many commercial laboratories, GC analysis is no longer an option. As a result, detection limits may be higher in order to satisfy laboratory logistics, sometimes sacrificing project requirements. The project team needs to be aware of these issues so they do not find themselves restricted by a method that does not fit project needs.

Often, calibration standards can be lowered in order to satisfy project objectives. Laboratories, however, often operate instrumentation within a specified calibration range. If a lower PQL is required, which may be closer to the MDL, the laboratory may be reluctant to alter their calibration standards. Generally, this is because laboratories need to be profitable and efficient and run several samples from many different clients using the same calibration setup. Lowering the calibration curve can often achieve a lower PQL, but may require additional dilution runs for more highly concentrated samples. Resolving these issues requires coordination with the laboratory, and may involve additional costs.

Example 3-16. Calculating the Method Detection Limit for Freon

GC/MS analysis of groundwater from a RCRA site following EPA SW-846 Method 8260B identified a Freon™ contaminant, 1-chloro-1,1,2-trifluoroethane. The concentration is very low or not detected in several of the wells at an MDL of 0.3 μg/L. The regulatory authorities would like to reanalyze the wells if a lower MDL and PQL can be achieved with associated precision (RSD) of less than ±15%. Investigation of this analyte by GC with an ELCD, in laboratory reagent blank water, provided a calibration as shown in the following figure.



Concentration versus response for the lowest four standard levels achieved the results shown in the following table. Analysis of laboratory blanks found no peak in this area of the chromatogram. Based on a comparison of the standard deviation to response, it appears that the results from the 0.02- and 0.03-μg/L runs have average responses that bracket 3 times their standard deviations. For this reason, the MDL is between 0.02 and 0.03 μg/L, and requires validation using a method such as 40 CFR Part 136, Appendix B.

concentration	.01			.02			.03			.05			
peak area	258	111	334	383	178	400	451	500	754	1178	1220	1089	1448
standard deviation	113			124			163			153			

™ Freon is a trademark of Dupont Corporation, U.S.

CHAPTER 4

DQIs RELATED TO ENVIRONMENTAL SAMPLING

4.1 REPRESENTATIVENESS

Definition and Concept

Representativeness was established as a DQI as a result of the recognition that characteristics of interest to environmental problems are heterogeneously distributed in space and time within the environment, and that careful attention must be paid during planning and implementation of a study to ensure that a set of samples adequately mirrors or reflects the characteristics of interest. Project managers and decision makers need assurance that the results of a particular data collection effort are not biased in any known way by the sampling or analysis design. Such a bias, combined with the inherent heterogeneity in the environment and uncertainty in the measurement process, could result in an incorrect conclusion. Lack of representativeness can have a direct impact on the ability to make the correct decision when relying on environmental data. To ensure representativeness, careful attention during the entire life cycle of a project is required.

The precise definition of representativeness is not straightforward; however, the American National Standard ANSI/ASQC E4-1994: *Specifications and Guidelines for Quality Systems for Environmental Data and Environmental Technology Programs* (ANSI/ASQC, 1994), offers the most relevant definition of representativeness for environmental studies and is adopted in this guidance document.

> "The measure of the degree to which data accurately and precisely represent a characteristic of a population, parameter variations at a sampling point, a process condition, or an environmental condition."

This definition of representativeness encompasses issues at both the micro- and macro-scale by addressing both how well measurements taken within a sampling unit reflect that unit ("parameter variations at a sampling point") and the degree to which measurements from a set of sampling units represent (allow you to make inferences about) the population of interest ("accurately and precisely represent a characteristic of a population"). Earlier sections on precision and bias indicators have already covered in detail some of the more quantitative elements of representativeness. Central to representativeness is assurance that both the sampling and measurement processes are free from known biases.

Representativeness is widely used in a less precise manner in the environmental community. For example, representativeness is a word commonly used to mean:

- there is an absence of biasing forces,
- it is a miniature or replica of the population,
- it is a typical or ideal case,
- there is a wide coverage of a population,
- it permits good estimation,
- it is good enough for the purposes of the study, or
- a statistically based sampling method was used.

While a number of these are indeed characteristics of a representative study, a more precise definition is needed to define suitable indicators of representativeness.

The word representativeness is called out in a number of EPA regulatory programs. For example, RCRA talks about " ... expected to exhibit the average properties of the universe or whole" (RCRA CFR 260.10). TSCA mentions "...at locations representative of the air entering the abatement site" (TSCA 40 CFR 763). In the air programs we see a discussion of the need for representativeness "...should be selected on the basis of spatial and temporal representativeness" (40 CFR 51-Appendix W), and the water programs simply state "...samples should be representative of daily operations" [40 CFR 403.12(b)]. While these statements lack a rigorous definition, they provide some understanding of the importance of this indicator to EPA programs.

The process involved in obtaining representative samples includes planning, implementation, and assessment. In addition to the sample design, careful attention must be paid to the measurement and analysis processes. Representativeness also has relevance in the world of laboratory research studies and experimental design beyond the scope of this document. For a discussion of these issues, the reader is referred to texts on experimental design, including Box and Hunter (1978) and Cochran and Cox (1957). Representativeness, as a DQI, is most relevant when viewed in the context of data's intended use. Several examples are presented to illustrate this concept.

Example 4-1. Basic Simple Random Sample for Representativeness

A site of interest is comprised of approximately 100 ponds that were used for settling of liquid wastes. In order to characterize the site, a statistical sampling design was implemented that called for one composite sample from three grabs near the center of each of 40 of the ponds. For this study, the sampling unit was defined as a single pond. The concentration of potential contaminants was expected to be homogenous across each pond, while differences were anticipated between ponds.

QC data were collected and analyses showed that the data were valid, with no bias detected. A data quality assessment was conducted verifying that the collected data met the site DQOs.

This data was considered representative of the site, and suitable for decision making.

4.1.1 Representativeness and Sampling

Consider the case where data are being collected to estimate the concentration of contaminants within some media (e.g., to define the nature and extent of a problem). Consider further that there is an expressed desire to generate a data set that reflects the variability and distribution of contaminants (or some other characteristic of interest) in the population of interest. This problem can be viewed at different scales. For example, the scale of the entire area of interest (macro scale), and/or the scale at which samples or measurements are made (micro scale). Representativeness addresses both the degree to which measurements "truly" reflect the concentration measured within the identified sampling unit, and the degree to which sampling units selected for sampling reflect the overall population of interest. As such, representativeness requires that the problem of selecting specimens from within the sampling unit, carefully defining the size and volume of each sample, and deciding how to measure the characteristic(s) of interest each be considered in order to obtain a precise and unbiased estimate. Representativeness also requires the problem of the number and location of sampling units that must be characterized to adequately reflect the distribution of the characteristic be considered, to support the intended use of the data set.

Figure 4-1 summarizes two alternative approaches to achieving a representative study. The left side of the figure addresses the classical statistical survey approach, and the right side addresses an alternative study design that is frequently encountered in environmental programs. The statistical survey approach involves probabilistic study designs intended to obtain an adequately representative sample from the population of interest. The alternative (depicted on the right side of Figure 4-1) involves studies designed to evaluate how experts believe that contaminants are being redistributed in the environment. The alternative is not strictly based on a classical survey design.

The pathway on the left side illustrates the steps leading from a probability-based design to the process of drawing inferences to the population of interest. Each scale of the problem is recognized, starting with the population of interest, then the population of accessible sampling units, and finally the sampling units themselves. Note that a probabilistic study design often allows one to draw conclusions directly about the population of interest. In some cases, it stops short because the accessible population is different from the population of interest (e.g., it is not possible to obtain a sample from all sampling units within the population due to some logistical constraints), and a judgment call must be made as to how well the accessible population represents the population of interest.

Given that it is not always practical or efficient to attempt to collect data from the entire population of interest (even using an efficient statistical design) to obtain an empirical representation of contaminant distributions, some studies are designed instead to verify a conceptual site model. These studies are designed in such a way that conclusions about the entire population can be made by linking the results of the individual studies together through a logic flow. For example, *if* data are collected from areas (or timeframes) that the model predicts

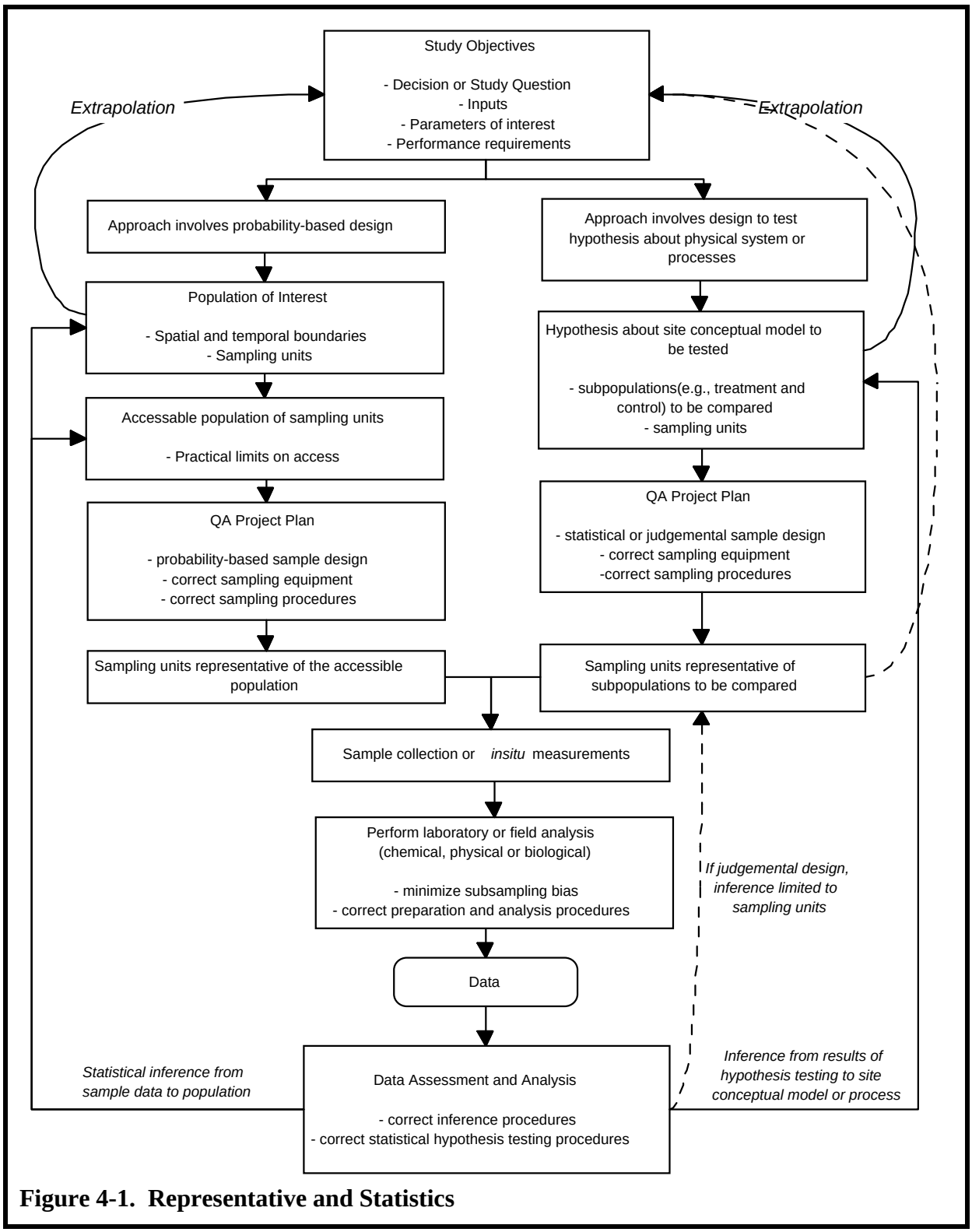


Figure 4-1. Representative and Statistics

will contain the highest level of contamination, *and* these data show that the concentrations are not above levels of concern (e.g., do not exceed regulatory thresholds, *or* do not pose an unacceptable risk), *then* the a decision can be made without performing an extensive study to empirically represent the entire population. This approach is commonly put to practice in sampling waste water discharges for compliance – the design does not attempt to represent all effluent; rather it focuses on specific periods of time when the levels of contamination are expected (based on process knowledge) to be at their highest. It is also frequently used to perform a screening assessment of a possible hazardous waste site.

A conceptual model can be developed to explain how a project team believes the processes related to the study topic operate. For example, a conceptual model may describe how contaminants have entered and become redistributed in the environment, or explain the temporal and geographic patterns of butterfly migrations, or the dispersal mechanism for broad-use pesticide applications. The conceptual model should reflect scientific judgment and historic knowledge of the problem. A data collection effort should be able to support or refute specific hypotheses regarding the conceptual model. The data collected might not be equally representative of all portions of the population, but rather provide the information necessary to either support or revise the current conceptual model.

The pathway on the right side of Figure 4-1 uses this alternative procedure to drawing inferences about a population of interest. In this case, the questions posed to verify the site conceptual model and subpopulations sampled to answer these questions impact representativeness. When a study has been designed to verify a site conceptual model, there is usually no attempt to ensure that every sampling unit within the population has some probability of being sampled. Smaller subpopulations (areas, time periods, or volumes) are frequently identified within the overall population of interest, and samples are taken to permit comparison among them. Either a judgmentally based sampling design can be used, or a statistical design that addresses the number of samples required in each area to see meaningful differences can be used. In either case, a greater degree of scientific judgment is required when attempting to draw inferences back to the population of interest, since these inferences will be based on the outcome of hypothesis tests and the associated logic flow, and not on a deliberate representation of sampling units within the population. Dashed lines represent a judgmental line of inference, wherein data represent sampling units, and inferences beyond the units are based on judgment.

In the conceptual model-related case, representativeness may take on slightly different dimensions. A representative study might be one where the outcome of the hypothesis test(s) can be used to make valid inferences about the study topic. The planning and design of such a study may still require a statistical approach. For example, statistics may be used to determine the number of samples required in areas expected to be impacted (i.e., treatment), and non-impacted (i.e., control), to detect meaningful differences with adequate statistical power. Allocation of samples within these subpopulations could then be addressed in a manner similar to the typical survey design; however, in most cases the boundaries of these populations will be greatly narrowed, based on the assumptions founded from the conceptual model.

4.1.2 Between-Sampling-Unit Representativeness

Define the population of interest. At the macro-scale, representativeness addresses how well the sampling units selected by the sample design represent the full population of sampling units available for observation (the sampled population), and in turn, how well the sampled population allows inferences to be made about the entire population of interest (target population). Failure to collect samples that fully reflect the population of interest, and failure to obtain precise, unbiased, measurements on these samples, may result in data that under- or overestimate the parameter of interest, resulting in the potential for a decision error.

Develop a statistical sampling plan. Statistical sampling theory addresses the range of approaches for selecting sampling units to support valid inferences about the population of interest. These approaches require that each sampling unit in the population have some known probability of selection. In general, the more variable units are within the population, the harder it will be to achieve a representative sample, and the larger the number of samples needed to support decisions regarding that population.

Statistical designs for environmental studies are frequently complicated by the non-random nature of the contaminant under investigation that results in correlation patterns between sampling units so that samples closer to one another are less different than those further apart. In these cases, samples close together do not meet the requirements of "independence" and design equations must account for this or risk underestimating the number of samples required to achieve representativeness. The more pronounced the correlation pattern, the more complex the required sampling design. To complete the statistical design process, some understanding of the conceptual model underlying the problem and the scale at which correlation is likely to exist must be obtained. Relevant estimates of variance are needed at both the within- and among-unit scales. This later requirement is frequently a limiting one, and often leads to a detailed analysis of historical data, and/or the design and implementation of a pilot study to support the design (sample size calculation process) of the full study. See *EPA Guidance for Choosing a Sampling Design for Environmental Data Collection* (U.S. EPA, 2000e) for further discussion on sampling designs.

Evaluate process for drawing inferences from data. The statistical design should have been optimized to support the intended use of the data. The sample size and allocation scheme required to represent the population of interest, or to perform statistical hypotheses tests about subpopulations of interest, should have been determined. Probability-based sampling designs provide the ability to evaluate the degree to which data are statistically representative of sampling units and the overall population of interest. In contrast, more professional judgment is required to evaluate the representativeness of other types of study designs, where not all sampling units have a known probability of being sampled.

4.1.3 Within-Sampling-Unit Representativeness

At the micro-scale, representativeness addresses how well actual physical specimens and the measurements performed on them reflect the true conditions within the defined sampling unit. For some media, such as water, obtaining a representative subsample is generally relatively easy; however, with particulate materials such as soils or sediment, this is often difficult. The importance of sample acquisition, homogenization, and protocols for subsampling is frequently underestimated. Gy's sampling theory stresses the importance of carefully determining the correct scale at which to sample particulate media, often referred to as the "correct module of observation" (Pitard, 1993). This task generally requires the results of a carefully designed pilot study to understand some basic, yet quite detailed, questions regarding the nature of the heterogeneity within the dimensions of the area under investigation. For example, if contaminants are highly clustered (patchy) on a scale much smaller than the scale of real concern, small grab samples may reveal quite varied results (some much above and some much below a level of concern), potentially skewing any distribution formed from a population of such values. Similarly, if the clustering behavior is not diminished by homogenization and subsampling procedures, representativeness will not be achievable, especially when the goal is to measure contaminants in a matrix with a large range of particle sizes. The procedure by which a specimen is obtained is also important. Care should be taken to ensure that the sampling device does not alter the proportion of each particle size that exists in the matrix under investigation.

Understanding the structural properties of the pollutants of interest allows a sampling protocol to be selected that addresses the required sample and subsample weights, and number of increments per sample needed to represent the characteristic of interest. Extensions of this theory may be applied to sampling of groundwater or other media. The theoretical underpinnings of Gy's work are complex, but the resultant conclusion is clear: many particulate media sampling methods violate theoretical principles of sampling. According to Gy, further research is needed in order to improve our ability to represent the true nature of contaminants in the environment.

Some specific ways to ensure within-unit representativeness include:

Use of correct sampling procedures and equipment. To achieve representativeness in environmental sampling, one must consider the potential for the actual sample acquisition process to distort the physical sample in a manner that may not provide material for analysis that mirrors the material present in the environment. For example, the collection of a particulate matrix such as soil can be affected by the choice of sampling device – a curved spoon versus a flat spoon (Pitard, 1993). The importance of carefully considering the process used to physically obtain and extract a specimen to achieve a representative, non-distorted sample from within a sampling unit should not be overlooked.

QA and QC requirements to ensure sample integrity. Once a physical sample is collected, a whole sequence of activities must take place prior to the sample actually being prepared for analysis in a laboratory. At each step in the sequence, the potential

exists for the constituents or characteristics of interest within that sample to be affected. Each step of the process where error could enter must be considered, and actions taken to prevent these errors. Following are examples of items or procedures that can impact sample integrity. Procedures to ensure integrity that are generally addressed in the QA Project Plan include:

- selection of sample collection jars,
- preservation of sample material,
- chain-of-custody,
- shipping conditions,
- sample receiving, and
- sample storage procedures.

Proper procedures will prevent alteration of the samples, and ensure that samples that reach the laboratory are similar to those in the environment, thereby preventing bias. For further guidance on ensuring sample integrity, see *EPA Guidance for Quality Assurance Project Plans*, Sections B1-3, B5, and B6 (U.S. EPA, 1998a).

Collection of an adequate amount of material. The volume, dimensions, and orientation of a physical sample from a solid medium such as soil or sediment can affect the ability to make inferences and therefore should be considered. The amount and weight of a sample should be considered in relation to the media particle size and shape (Pitard, 1993; Lame and Defize, 1993). In addition, it is important to consider the amount of material needed by the laboratory to perform the desired suite of analyses. All inputs to the decision (DQO Step 3, EPA QA/G-4), (U.S. EPA, 1994) must be identified prior to determining the volume of material that must be obtained in the sampling effort. It is also important to consider what portion of the media inferences will be made about, to ensure that the sampling protocol is well suited for representing this portion. For example, if the primary concern relates to the very fine grained soil particles that might adhere to the skin and be incidentally ingested, or to the fine particulates in air less than a few microns in size, the sampling protocols must be carefully designed to obtain samples that represent these size fractions in proportion to the total population of interest.

Compositing is frequently used to increase sample representativeness by creating a physical average of sample increments collected over some area for which the average properties are of interest. Composites can be comprised of multiple grabs from within a sampling unit, or can be formed from grabs taken from multiple sampling units, depending on the goal of a particular study. If within-unit composites are formed, this approach has the effect of increasing sample support, assuming that complete homogenization and representative subsampling is achievable. When compositing over small or large areas, a grid or systematic sampling scheme or a random sampling scheme can be used to ensure that the grabs taken to form the composite represent the area of interest.

Compositing strategies can greatly improve the representativeness of a set of samples, if the goal of the study is to estimate the average properties over some set area. For additional information on the incorporation of compositing strategies into the sampling design, the reader should consult *EPA Guidance for Choosing a Sampling Design* (U.S. EPA, 2000e), or Lancaster and Keller-McNulty (1998).

Subsampling in the laboratory is generally required to prepare samples for analyses. To minimize bias that could be potentially introduced with this procedure, it is critical to carefully follow procedures for homogenization, and if a particulate matrix is being sampled, to perform "correct sampling procedures," in a manner similar to obtaining the field sample. Pitard (1993) describes such procedures that are aimed at avoiding the introduction of bias through the use of sampling procedures that result in a different assemblage of particle sizes in the materials to be prepared than are in the field.

Selection and implementation of an analytical measurement method, including sample preparation (extraction, cleanup). When selecting measurement methods to ensure that the measurements adequately reflect the true conditions in the environment, considerations such as the sample preparation and extraction procedures must be considered. These considerations can affect the representativeness of a set of samples as much as the sample design itself. For example, when one is trying to determine the potential for hazardous waste to leach into groundwater, a weak acid leach procedure, known as the toxicity characterization leaching procedure is usually recommended. This procedure mimics the acidity that can build up in a landfill, and thereby represents what might be expected to occur to wastes placed in that landfill over time. When looking at the potential human health risk, samples are typically prepared using a partial acid digestion (weak acid leach). Again, this procedure mimics the acidity of the human digestive system and provides a representation of the contaminants that might become available (and therefore taken up by the body) if ingested, rather than representing the actual chemical make up of the parent material (sediment, soil, etc.) that you would represent by preparing the sample with a hydrofluoric acid, total dissolution of the material. These issues are raised to emphasize the importance of discussing sample preparation and analytical methodology with the project chemist to ensure that analytical results will be appropriate for the intended use.

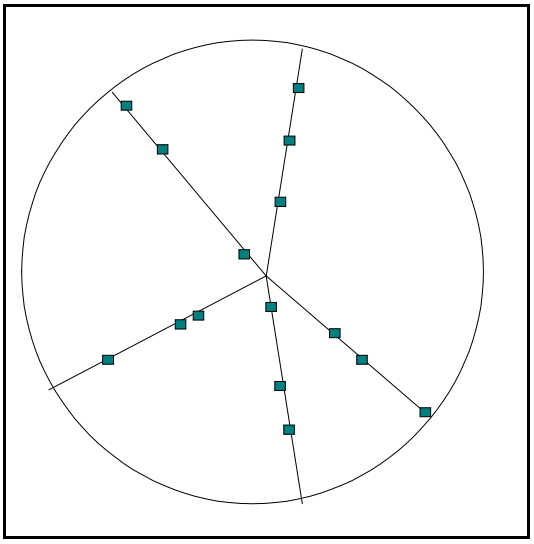
4.1.4 Assessing Representativeness

In the environmental sciences, when the question is asked whether an existing data set can be used for a purpose other than originally intended, a determination of whether the data set is representative of the population of interest for this new use is necessary. The checklist described in Table 4-1 can be used to evaluate representativeness. It may also be useful when designing a new study to ensure representativeness. The basic questions to be answered are whether the individual measurements of the characteristics of interest accurately reflect the conditions in the sampling unit, and whether an adequate number of units were measured to

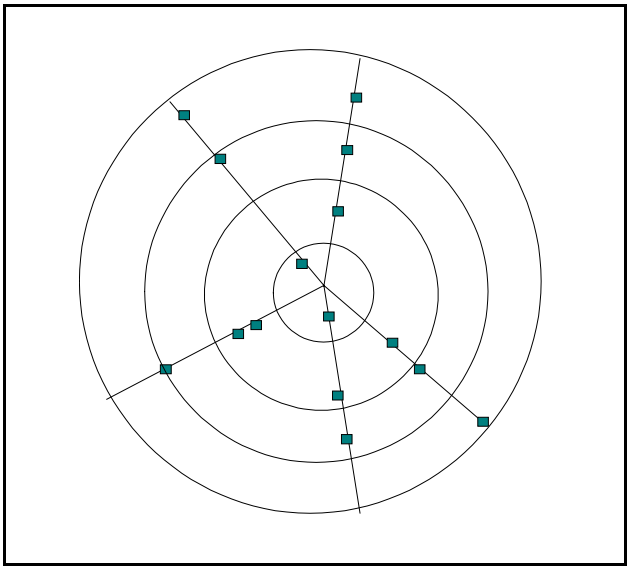
Example 4-2. Post-Hoc Weighting for Representativeness

Characterization data were collected from the locations illustrated on the following schematic of the area around a paper manufacturing plant. Five spokes from the emissions stack were selected by a random number generator based on the angle from a randomly set starting direction. (Spokes were selected randomly due to marked fluctuations in seasonal wind patterns.) On each spoke, three sampling locations were selected by random number generator based on their distance from the stack. These data were analyzed for the chemicals of interest, and the data are now available for statistical analysis. Are the selected sampling units spatially representative of the population (area impacted by stack emissions) of potential sampling units?

Sampling Locations for Characterization



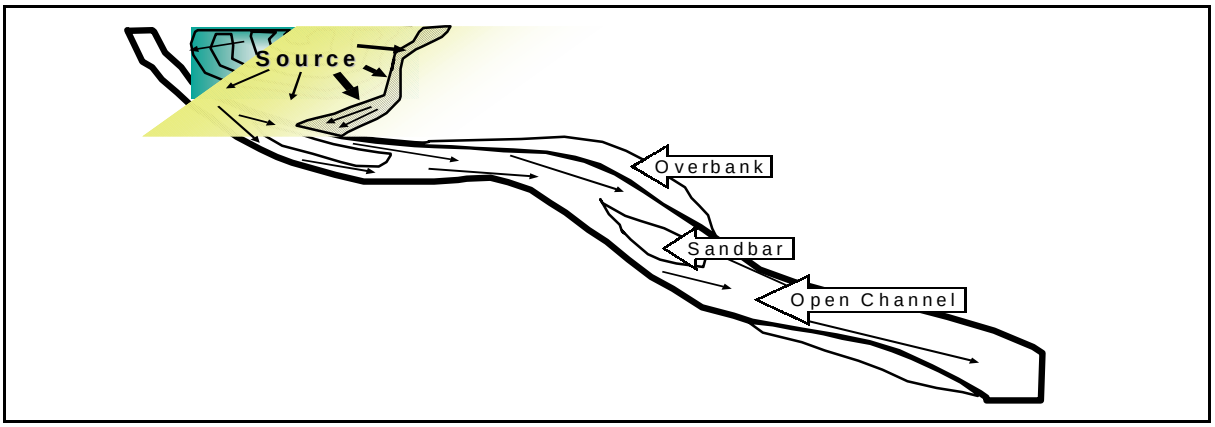
Because this random allocation was done on a circle, with linear random number generation, the data under-represent the area near the outer rim of the circle. For an accurate assessment of the population, it is necessary to understand how the data represent the population. In this case, the statistician and project team used a post-stratification of the area of the site to make the available data applicable for site-wide estimates of chemical concentrations. This was accomplished by laying four concentric rings over the figure of the site such that each of the rings was of equal width, as shown in the following figure.



The area of each ring was calculated, and the samples within each ring were combined according to the relative area of the site represented by their ring. Through the use of this post-hoc method for weighting the data according to the relative area each sample represents, the project team was able to make decisions

Example 4-3. Conceptual Model Driven Design

Consider a watershed system comprised of an intermittent stream bed, some 20 miles in length, where the risk associated with some upgradient source is in question. A physical, conceptual model can be constructed to lay out how scientists familiar with the behavior of the stream system in general believe the chemicals of concern will behave in the system. To gather an adequate number of samples to support decisions about the system, a standard probabilistic design could be developed using relevant estimates of variance, and determining the number and location of samples needed to support an assessment of risk. Alternatively, one could conduct a series of studies to test hypotheses about components of the system and determine if the overall population of interest can be narrowed down to some smaller subset, and a decision made later as to whether additional characterization of this narrowed population is required.



In this case, scientists reason that the natural processes in the watershed have redistributed the source material in a predictable manner – higher levels of metals adhered to finer particles, which in turn have been deposited in overbank, sandbar, and channel areas differentially. In addition, one could reason that concentrations within these features would diminish with distance from the source term. If these types of questions can be tested empirically, then the study can focus on certain geomorphologic features and in exclusion of others, greatly decreasing the demand for data to fully characterize the entire system. The approach to "representing" the system adequately to support decisions is one involving a logic flow leading from results of hypothesis tests to conclusions about

reflect the population of interest. Sample support (the finite volume of the specimen that is measured to provide a single observation) may affect the scale to which the observation can be used to draw inferences, since what is seen often depends on the scale observed. In answering questions about within-unit representativeness, a review of the results of quality assessment samples such as field duplicates (collocated samples), splits, or other replicates should be performed. In addition, information on laboratory performance (precision, bias, and sensitivity) is needed. Finally, care should be taken to compare the scale of the measured subsample with the scale of the sampling unit, and procedures that were taken to obtain the specimen for analysis to ascertain whether the result is likely to reflect the properties of the unit. Finally, procedures used

to homogenize the specimen and subsample for analysis should also be considered to be sure that adequate procedures to control within-specimen variability were followed. Assuming that data are representative, at least to some degree, of the population of interest, EPA QA/G-9 (U.S. EPA, 1996) provides useful guidance on determining whether data meet the user's DQOs.

Table 4-1. Checklist Form of the Representativeness DQI

Attribute of Representativeness	Importance of Attribute
Were study objectives adequately defined using the DQO process or its equivalent?	Representativeness is only meaningful in the context of the intended use of data.
Was the population of interest unambiguously defined to include spatial and temporal boundaries and a thoughtful definition of sampling units comprising the population?	Defining the population as a set of sampling units is a minimum requirement for developing a probability-based design. Selection of the dimensions of the sampling unit can help the investigator move away from drawing conclusions concerning an entire site based on observing tiny areas of contamination, and may help avoid problems of samples not being independent.
Was the statistical basis for the sampling plan explained, including the basis for determining the number and allocation of samples within the population of interest?	Representativeness hinges on an adequate number of samples being collected. Different schemes for allocating samples to maximize their effectiveness can be used. Sample design theory provides a range of approaches that are adaptable to environmental studies.
Was a rationale provided to support the selection of sampling equipment and procedures to ensure samples sent to the laboratory mirror what is in the environment?	Correct choice of equipment and procedures can make the difference between obtaining samples that reflect the true characteristics of the matrix within the sampling unit, or obtaining a sample that is not reflective of the sampling unit because of the collection equipment and method used.
Was the rationale provided for the selection of analytical methods, including sample preparation and extraction, and are these methods appropriate for generating the measurements needed to support the study objectives?	Results from analytical instruments or other types of laboratory tests (e.g., bioassays) will only be useful in drawing inferences if the measurements reflect the characteristics of true interest to the investigator.
Were samples collected from all selected sampling units following quality procedures spelled out in the project QA Project Plan?	Incomplete sampling, if biased, can lead to spurious conclusions about the population.

Table 4-1. Checklist Form of the Representativeness DQI

Attribute of Representativeness	Importance of Attribute
Were QA and QC steps followed that were designed to ensure sample integrity?	Loss of sample integrity that may occur if quality procedures are not followed can result in loss of analytes, and hence bias in the results. Biased results will not be representative.
Were analytical procedures followed, and MQOs for precision, accuracy, and sensitivity achieved?	Within-sampling-unit sources of error must be controlled. This control may be limited to subsampling and analytical error, when the sampling unit is defined as the portion of media sampled, or may include small-scale variability, when the unit dimensions are larger than the physical sample.
Were appropriate methods for data assessment and analysis performed to draw inferences or perform hypothesis tests?	Appropriate methods for estimating parameters, including the variance, follow from the choice of sample design in order to obtain values representative of the population of interest.

4.2 COMPLETENESS

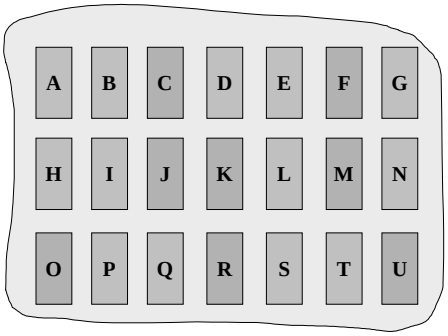
Completeness is a measure of the amount of valid data obtained from a measurement system, expressed as a percentage of the number of valid measurements that should have been collected according to the study design (i.e., measurements that were planned to be collected). Percent of completeness is calculated using the following formula:

$$\text{Percent Completeness} = \frac{\text{(number of valid measurements) x 100}}{\text{(total number of measurements planned).}}$$

Completeness is a measure of how well a sampling and analysis design was implemented. A data set that is 100% complete is the result of careful and precise implementation of the data collection plan. Completeness is not intended to be a measure of representativeness; for example, it does not describe how closely the measured results reflect the actual concentration or distribution of the pollutant in the media sampled, but may be a contributing factor. A complete data set may or may not achieve the DQOs depending on how well the sampling plan reflects the conceptual model for the site, how accurate the conceptual model was prior to sampling, and whether the distribution of the reported data results are similar to the anticipated results.

Example 4-4. Using a Pilot Study to Achieve Representativeness

Characterization of a site containing 21 evaporation ponds was initiated with very little information on the possible sources of contamination, methods of contaminant dispersion, or anticipated levels of contaminant concentrations. The goal was to determine whether remediation would be required based on human health and ecological risk thresholds. DQOs for the project were determined, but sufficient funds were not available to meet the DQOs based on the uncertainties about the site. Hence, a pilot study was designed to gather information on which a more complete and efficient sampling and analysis plan could be based.



Eight ponds were randomly selected for the pilot study. Six samples from random locations in each of those ponds were collected. The arsenic results are summarized in the table below.

Arsenic	**M**	**R**	**C**	**O**	**F**	**J**	**K**	**U**	**Overall**
Mean (in ppm)	20.3	26.5	30	21.8	19.3	15.7	15.1	18.2	20.9
Variance	657	117	56	122	368	144	236	264	233

A statistical test conducted on this data clearly showed that for the existing data, there is great similarity between ponds. In fact, the contribution to total study error is greater from the within-pond component than from between ponds. Similar tests were performed on all the different analytes for which data were collected in the pilot study. The finding of no significant differences between ponds was consistent across the analytes reported.

	Sum of Squared Distances from Mean	Percent Contribution
Between Ponds	1,114	10 %
Within Ponds	9,816	90 %
Total	10,930	100 %

Armed with this information, the project team chose to collect composite samples from within each pond. A few composite samples from each pond was considered adequately representative of the site based on the information gained in the pilot study. These composite analyses will decrease the variability of data within each pond so that comparisons of each pond to criteria of interest may be made with sufficient power for decision making within the budgetary constraints of the project.

Example 4-5. Quantitative Measure of Completeness Doesn't Tell the Whole Story

In the early phases of environmental restoration at one of the nation's nuclear laboratories, data were collected from a large area of the site according to a very thorough systematic sampling grid. These data were analyzed as requested and used for decision making for each of the solid waste management units within that area. Unfortunately, as a result of some confusion over the way the request for metals analyses was expressed, no results for mercury were provided. Because so many other analyses were requested, and the chemical laboratory did a fine job of completing valid analyses on most everything other than Hg, the% completeness of that data set was between 90 and 95%. While this may sound like an adequate percentage of valid results, no conclusions could be made about any solid waste management units for which mercury was a potential concern.

4.2.1 Does Data Need to be 100% Complete to be Useful?

The important question for decision makers is whether the number of measurements is sufficient to support the decision to be made. For example, there could be only 70% data completeness (30% lost or found invalid), but, because of the nature of the study design, the results could still be representative of the target population and yield valid estimates. Conversely, a data set with much higher completeness, but systematically omitted or rejected data may be insufficient to yield valid estimates of the parameters of interest.

When evaluating completeness, it is useful to consider not only whether an acceptable percentage of samples from the total planned number was collected, but also whether an acceptable percentage of analytical results was obtained. A simple process for evaluating completeness of a data set involves preparing and evaluating a data summary table. The data summary table indicates if valid measurements for each analyte or other variable of interest were obtained at each sample location. Incomplete information will be evident during review of data summaries of this type.

Another aspect of completeness involves evaluation of the data package to determine if all associated QC data and required calculations (such as detection limits) were provided. The process of checking for the completeness of a data set, sometimes termed verification, is routinely done upon receipt of a data set, prior to approving payment for the analytical work from a laboratory.

4.2.2 What is the Effect of Incomplete Data?

The degree to which lack of completeness affects the outcome of a study is a function of many variables ranging from deficiencies in the number of field samples acquired to a failure to analyze as many replications as deemed necessary by the QA Project Plan and DQOs. While the percent completeness of a data set can be expressed quantitatively, the intensity of the effect caused by incompleteness of data is best expressed as a qualitative measure.

One method that may be employed to minimize the chance of having insufficient data for decision making is to increase the requested samples and analyses beyond the number of

measurements estimated to demonstrate compliance to account for lost or rejected data and uncertainty in the calculation of the number of measurements. For example, if the number of measurements estimated was increased by 20%, this means that a study with only 83 percent completeness may still have sufficient power to support the decision of interest. However, this method can be costly (in this example, the sampling and analytic costs are increased by 20%) and, depending on which data are lost, rejected, or not collected, sufficient data for decision making still may not be obtained. Constructing a statistical power curve and evaluating the results will help determine if the number of measurements is sufficient to support the decision. This approach is most beneficial in situations where there is previous information on the rate of lost or rejected analytic results from the laboratory, where initial sampling and analysis costs are not overly constraining, and when it is necessary to make decisions on only one set of data because of time constraints.

Completeness can have an effect on DQO parameters. Lack of completeness may require reconsideration of the limits for the false acceptance and false rejection rates because insufficient completeness will decrease the power of the statistical test. If the degree of completeness is a cause for failing the DQO requirements, expert opinion may then be required to ascertain if further samples are necessary.

Lack of completeness is often a vital concern with stratified sampling. Substantial incomplete sampling of one or more strata can seriously compromise the validity of conclusions from the study. In other situations (for example, simple random sampling of a relatively homogeneous medium), lack of completeness may result only in a loss of statistical power, which may, in turn, result in a failure to meet the desired DQOs.

4.2.3 What are the Causes of Incomplete Data?

Several factors may result in lack of completeness, such as:

- the DQOs may have been based on poor assumptions,
- the study design may have been poorly implemented,
- the design may have proven impossible to carry out given practical constraints or resource limitations, and
- the requirements of the QA Project Plan may not have been fulfilled.

Lack of completeness should always be investigated, and the lessons learned from conducting the study should be incorporated into the planning of future studies. Table 4-2 presents the minimum considerations, impacts, and corrective actions for completeness.

Example 4-7. Sample Data Summary Table for Completeness

This sample data summary table illustrates an easy, tabular format for checking that the requested samples were collected, the appropriate analyses were conducted on each sample, and the data are valid.

Sample ID	Location ID		Depth (Inches)		Analyses									
					SVOCs		VOCs		Metals		TOC		% Moisture	
98-003-01	32x47y	✓	0-6	✓	Y	✓		✓	Y	✓	Y	✓	Y	R
98-003-02	32x47y	✓	12-18	✓	Y	✓	Y	✓	Y	✓	Y	✓	Y	R
98-003-03	32x47y	✓	24-30	✓	Y	✓	Y	✓	Y	✓	Y	✓	Y	R
98-003-04	18x29y	18x3	0-6	✓	Y	✓		✓	Y	✓	Y	✓	Y	R
98-003-05	18x29y	18x3	12-18	✓	Y	✓	Y	✓	Y	✓	Y	✓	Y	R
98-003-06	18x29y	18x3	24-30	✓	Y	✓	Y	✓	Y	✓	Y	✓	Y	R
98-003-07	14x22y	✓	0-6	✓	Y			✓	Y	✓	Y	✓	Y	R
98-003-08	14x22y	✓	12-18	✓	Y		Y	✓	Y	✓	Y	✓	Y	R

R = Rejected data
SVOC = Semivolatile organic compound
TOC = Total organic compounds
VOC = Volatile organic compound
Y = Analysis requested

In this example, several differences between the expected data and actual data were observed.

- The location from which three of the samples were collected is different than expected. A physical impediment to sampling at the proposed location had occurred, and an alternative location was selected in accordance with guidance written into the QA Project Plan for just such an event. These data are considered valid for calculation of completeness, but should be re-evaluated when considering the DQIs for representativeness.
- The laboratory ran VOC analyses on all samples, even those for which it was not requested. This procedure neither impacts completeness nor adversely affects data analysis.
- All percent moisture results are rejected. The lack of information on percent moisture may make some decision making more difficult or impossible based on this data.
- Some SVOC results are missing. Depending on the robustness of the DQOs that were the basis of the sampling plan, the observed results, and the relative importance of the locations from which SVOC results are missing, this lack of completeness could also adversely impact the ability to make decisions based on this data set.

The percent completeness of the portion of this data set illustrated in the above table is:

Table 4-2. Minimum Considerations for Completeness

Considerations for Completeness		Impact when Considerations are Not Met	Corrective Action
Sample Collection	Were the number of samples requested actually collected?	A reduction in power for decisionmaking may occur. If the missing samples are predominantly from one area, time frame, or strata, serious bias may occur.	Resurveying, resampling, or reanalysis to fill data gaps may be necessary. Additional analysis of samples already in the laboratory may be possible.
	Were samples collected from the locations or sources requested?	Serious compromising of study design if major deviation from planned sampling scheme. Possible bias if minor or systematic deviation from planned sampling scheme.	Resurveying, resampling, or reanalysis should be considered to fill data gaps.
	Were samples collected in the time frame requested?	Serious possibility of bias if the samples were not collected in time periods that reflect the target population.	The time frame of interest may have passed, or, if it is cyclic (e.g., seasonal), then data may be collected during next cycle.
Completeness	Were the number and type of analyses requested actually conducted?	Spatial and temporal aspects of missing analyses should be considered to determine if missing analyses cause bias as the result of an underrepresentation of a portion of the population of interest. Reduction in power for decision-making may occur.	Additional analysis of samples already in the laboratory may be possible. Resurveying, resampling, or reanalysis may be required to fill data gaps.
	Were the QC sample analyses requested conducted?	Validity of the data should be carefully considered because bias may result and decision making may be hampered by this bias.	Additional analysis of samples already in the laboratory may be possible. Attempts may be made to generate the missing data from existing information.
	Were any data rejected during the verification and validation phase?	Spatial and temporal aspects of rejected analyses should be considered to determine if missing analyses cause an underrepresentation of a portion of the population of interest. A reduction in power for decision making may occur.	Additional analysis of samples already in laboratory may be possible. Resurveying, resampling, or reanalysis may be required to fill data gaps.

CHAPTER 5

DATA QUALITY INDICATORS BEYOND PARCCS

A number of DQIs beyond the PARCCS terms are commonly used by various programs to describe or evaluate the quality of specific data sets. A brief introduction of the following additional DQIs is presented:

- Reproducibility
- Repeatability
- Integrity
- Validity

This section of the DQI guidance will continue to be augmented over time to capture and discuss useful indicators.

5.1 REPRODUCIBILITY AND REPEATABILITY

Reproducibility is a qualitative indicator that is used to refer to the uncertainty associated with the use of multiple laboratories for a specific study. Other intuitive and common uses of the term, such as to describe the ability for multiple researchers to arrive at the same conclusions, are beyond the scope of this guidance. The ability of multiple laboratories to generate the same result for splits of the same material can be expressed as a measure of interlaboratory precision and bias. Specific indicators of precision and bias (such as the range or variance) are simply generated using data sent to multiple laboratories.

Repeatability generally refers to the degree of agreement between independent test results produced by the same analyst using the sample test method and equipment on random aliquots of the same sample within a short time period. Mandel and Lashof (1987) have written extensively on these indicators. Both indicators were developed to discuss the results of interlaboratory studies of test methods. Repeatability was defined as the probability that two test results obtained in the same laboratory (on the same material) will not differ by more than some specified amount. Reproducibility refers to the same thing, except that it measures this probability for multiple laboratories (or operators, apparatus, and over time). To be viewed as repeatable or reproducible, the probability is set at 95%.

5.2 TECHNICAL INTEGRITY (includes traceability, completeness of records, chain of custody, etc.)

Technical integrity is a qualitative indicator that infers that a datum is valid and has not been compromised due to improper or inadequate handling, documentation, analysis, review, or archival. During the evolution of a data point, a variety of factors can adversely impact its

technical integrity. Examples of the kinds of questions that should be asked to ascertain integrity are listed below.

1. *Assessment of the sample collection and handling procedures*. Were the SOP-approved sampling procedures followed? If not, could the deviation from those procedures impact the quality of the sample? Were field conditions adequately characterized and documented? If not, does the absence of this information impact the ability of data analysts to interpret analytical values? Did previous audits, inspections, or surveillance activities reveal deficiencies that could impact the quality of the sample? Were sample holding times (i.e., the time between sample collection and sample analysis) exceeded in the field or during shipment? Was the sample properly preserved?

2. *Assessment of the analytical procedures*. Were sample holding times exceeded in the laboratory? Was each sample analyzed by a laboratory that was qualified to conduct the requested analyses? Had that laboratory been audited within the time agreed upon in its contract? For each laboratory, what fraction of the analyses were rejected during data validation?

3. *Assessment of the documentation procedures*. Are all records related to a sample traceable? Are records available to connect samples with known locations? Are complete Chain-of-Custody forms filed for all samples? Are the reported data present in the data bases developed to facilitate data analysis?

4. *Assessment of the previous technical and QA/QC reviews*. Were the data appropriately verified and validated? What fraction of the total data set was verified and validated and verified? What types of problems were identified in the verification and validation and verification process?

In some instances, determining data integrity is a subjective decision. Improper or inadequate handling, documentation, analysis, review, or archival may render a datum unusable, or just questionable. Questionable data may or may not be unusable.

Technical integrity is distinct from legal defensibility, although the two conditions are related. A condition that might cause a datum to be challenged in a court of law is not necessarily a condition that impacts technical integrity. For example, failure to initial changes in a field notebook may result in a legal challenge of the analytical data resulting from that field activity; however, it does not necessarily indicate that the data values are incorrect. Similarly, failure to maintain files of completed chain-of-custody forms does not necessarily mean that the related analytical data have been compromised.

To conduct a thorough evaluation of technical integrity, field records must be available for review. The critical field records are the field notebooks, sample collection forms, and chain-

of-custody forms. Reviewing analytical data and their management is another critical phase in a technical integrity review.

5.3 VALIDITY

"Validity" is a term often applied to data points or data sets. Validity is not a DQI *per se*, but it is related to the achievement of measurement quality objectives based on DQIs. Validity is typically determined by the completion of processes known as data verification and data validation. The EPA *Guidance on Environmental Data Verification and Data Validation* (EPA 2001a) defines these processes as follows:

> **Data Verification** is confirmation by examination and provision of objective evidence that specified requirements have been fulfilled. Data verification is the process of evaluating the completeness, correctness, and conformance/compliance of a specific data set against method, procedural, or contractual requirements.
>
> **Data Validation** is confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use are fulfilled. Data validation is an analyte- and sample-specific process that extends the evaluation of data beyond method, procedural, or contractual compliance (i.e., data verification) to determine the analytical quality of a specific data set.

It is clear from these definitions that the two processes are related. Nevertheless, there is a fundamental difference in their respective emphases. Data verification is primarily an evaluation of performance against pre-determined (and often generic) requirements given in a document such as an analytical method procedure or a contract; it addresses issues like those described in the Technical Integrity section above. Data validation, on the other hand, focuses on particular data needs for a project, which are generally expressed in a QA Project Plan in terms of one or more of the quantitative or qualitative DQIs discussed in earlier chapters of this guidance. Data points or data sets are described as "valid" if the data validation process has determined that they are consistent with project-specific DQI objectives. The data validator routinely qualifies "invalid" data by means of alphanumeric flags or narrative descriptions.

CHAPTER 6

DQIs IN THE PROJECT LIFE CYCLE

6.1 THE ROLE OF DQIs IN PROJECT PLANNING

DQIs play a number of important roles during the project planning phase, such as calculating relevant indicators from historical data to support the design of new efforts, establishing MQOs and identifying future DQI needs, and specifying quantitative and qualitative requirements in the QA Project Plan.

6.1.1 Historical Data Review

The role of DQIs during project planning focuses on assumptions (typically based on analysis of historical performance data) made during the development of DQOs and an associated statistical design of a project. For example, design optimization may involve an analysis of the major sources of sampling and measurement error, as well as spatial and temporal variability that will contribute to uncertainty. Total study error impacts the ability to achieve the limits on decision error specified during Step 6 of the DQO process (U.S. EPA, 2000d). Depending on the relative contribution of different components of total study error (especially components of total study precision), different choices may be made to cost effectively achieve the specified DQOs.

Tradeoffs can be made regarding selection of analytical methods, sampling methods, and the number, location, and timing of samples (or direct measurements) taken. For example, a common tradeoff is one between collecting a large number of samples and performing relatively inexpensive, rapid screening analysis (e.g., XRF for metals and immunoassays for PAHs and PCBs), versus collecting fewer samples and sending them to a fixed laboratory for full suites of metals, semivolatile organics and pesticides/PCBs. In both cases, a similar control on total variance may be achievable, however one option is usually more cost effective than the other. Typically, the fixed laboratory provides lower detection limits and more complete documentation, but given the relatively higher costs involved, fewer samples can be taken to represent an area. In making these important tradeoffs, assumptions are made regarding the expected analytical performance (e.g., precision, bias, sensitivity) of the various available measurement methods, given the specific matrix and measurements of interest. In addition, some assumptions must be made regarding the distribution and variability of the constituents of interest, as well as regarding the process(es) controlling their release and distribution within the environment. Ideally, these assumptions should be based on an analysis of historical data or on a pilot designed to obtain this information.

DQIs for precision, bias, and sensitivity are critical inputs to supporting the kinds of tradeoffs discussed above. Alternative designs are developed that reflect different combinations of statistical sampling and measurement schemes. In essence, these designs reflect the use of

DQIs for these different schemes. The expected performance of these designs is calculated using these indicators as inputs, but the output of this effort is only as sound as the assumptions (including the quality and relevance of the indicators) that are made. In addition, once the critical DQIs are identified (those that the final data quality are expected to be most sensitive to), important design choices concerning QA and QC samples required to enable estimates of the DQIs can be made.

6.1.2 DQIs as Inputs to the QA Project Plan

Decisions about what DQIs are needed in a study and what MQOs must be achieved to meet the project DQOs should be documented in the QA Project Plan. The structure of the QA Project Plan (Table 6-1) includes several sections where DQI information can be inserted:

- Section A7.2, *"Specifying Quality Objectives,"* the guidance on Quality Assurance Project Plans (U.S. EPA, 1998a) notes that "...DQIs can be evolved from DQOs for a sampling activity through the use of the DQO process. DQIs and associated MQOs should be considered during the design optimization process, and documented as part of the Quality Objectives..." in the QA Project Plan.

- Section A7.3, *"Specifying Measurement Performance Criteria,"* the guidance notes that one of the most important features of the QA Project Plan is that it links the user's quality objectives to verifiable measurement performance criteria. MQOs are statements of measurement performance criteria, linked to the design decisions made as part of the DQO process. When appropriate, these MQOs should be documented in this section of the QA Project Plan. This is especially true for situations where project-specific requirements are necessary (i.e., when laboratory or method default requirements are not sufficient, and project-specific requirements are established), or when performance-based methods are selected for use in a study.

- Section B1.4, *"Design Assumptions,"* the guidance recognizes that a number of assumptions are made during the design, upon which the success of the design may rest. It would therefore be logical to state what DQIs are needed to test these assumptions (such as testing the ability of the method to achieve required PQLs, precision requirements and recovery specifications, or evaluating the sampling method itself). If MQOs were specified that lend themselves to real-time evaluation in a QC mode, this explanation should state not only the assumptions, but also contingency plans, should the assumptions be found not to hold during a study.

- Section B5.2, *"Quality Procedures,"* the guidance discusses QA acceptance limits and project-specific requirements that may go beyond defaults. QC checks for the

field and laboratory, needed to determine if requirements are met, should be specified here. These checks will usually be comparisons of DQIs to MQOs; hence a direct connection exists between the DQI guidance provided in this document and the QA Project Plan. This section also calls for specifying the precise formulae to be used to calculate the required indicators.

Table 6-1. List of QA Project Plan Elements

Project Management	Measurement/ Data Acquisition	Assessment/Oversight
A.1 Title and Approval Sheet	B.1 Sampling Process Design (Experimental Design)	C.1 Assessments and Response Actions
A.2 Table of Contents	B.2 Sampling Methods	C.2 Reports to Management
A.3 Distribution List	B.3 Sample Handling and Custody	
A.4 Project/Task Organization	B.4 Analytical Methods	**Data Validation and Usability**
A.5 Problem Definition/Background	B.5 Quality Control	D.1 Data Review, Verification, and Validation
A.6 Project/Task Description	B.6 Instrument/ Equipment Testing, Inspection, and Maintenance	D.2 Verifications and Validation Methods
A.7 Quality Objectives and Criteria	B.7 Instrument/Equipment Calibration and Frequency	D.3 Reconciliation with User Requirements
A.8 Special Training Certification	B.8 Inspection/ Acceptance of Supplies and Consumables	
A.9 Documentation and Records	B.9 Nondirect Measurements	
	B.10 Data Management	

- Section B9.2, *"Acquisition of Non-Direct Measurement Data,"* the guidance discusses a number of DQIs and how they pertain to non-direct measurement data (secondary data). DQIs are necessary to aid in determining if the use/reuse of a data set is appropriate. It is important to determine what data quality information should travel with a data set to help current as well as future users characterize the quality of that data set.

- Appendix D of the guidance provides a brief overview of each of the principal DQIs covered in this guide, plus several others that are either not addressed here or are subsumed into the principal DQIs (for example, recovery is discussed here under accuracy or bias). Section AD2.7 specifies that at a minimum, the QA Project Plan must address precision, accuracy, representativeness, comparability, and completeness, and provides a table that discusses the sources of errors associated with these indicators.

6.2 DQIs IN PROJECT IMPLEMENTATION

By continually evaluating the performance of the measurement system, the need for adjustments or corrective action to keep measurement systems in control can be identified before the project budget and schedule are expended. This is especially true for ongoing measurement programs, such as monitoring efforts required for compliance programs, because mid-course corrections can be made based on feedback from the last sampling effort, or last series of samples. To support these QC measures, specific DQIs must be identified, and data collected to support their calculation. In addition, MQOs are needed to assess and interpret the QC data results. Based on these types of comparisons, corrective actions can be taken as needed to improve project performance and increase the probability that data will be adequate to support the intended use.

In programs where large numbers of samples are being collected and multiple laboratories are performing the analytical work, DQIs can be calculated using carefully designed QC samples. Questions such as, "how well are the various laboratories performing?," "are critical indicators such as sensitivity, precision or bias (frequently estimated as recovery) being met?," or "are changes to the analytical protocol (such as sample preparation) needed?," can be addressed periodically, thus providing an opportunity to make changes needed to generate an acceptable data set. Answering these questions is part of Statistical Process Control, a topic beyond the scope of this guidance.

To take full advantage of the careful design work, a viable QA oversight effort is required. Too often, QC data are generated, but not evaluated at all, or not evaluated in a timely, meaningful way. To assist EPA managers in understanding how to build quality into their efforts, and the role that DQIs play, the DQIs discussed in this document can be used not only in a real-time QC mode, but also in the assessment of the overall performance of a program over some period of time, such as a field season. For example, indicators of precision, accuracy, and sensitivity can be evaluated over the field season to determine how well the program was able to measure critical variables. In addition, comparisons can be made between laboratories to determine if any of them are routinely failing to provide adequate data, to identify problems that must be resolved, or to ensure comparability when data are to be pooled.

6.3 DQIs IN DATA ASSESSMENT AND REPORTING

Following a data collection effort, a determination of the adequacy of the data can be made following the Data Quality Assessment Process as described in EPA QA/G-9 (U.S. EPA, 1996). At this point, the highest interest is in whether the data set will support a decision with the desired degree of certainty. It is important to consider the performance and representativeness of the measurement effort prior to reaching conclusions regarding data adequacy; however, at this point it is less critical to determine if each and every goal set for given DQIs (i.e., the MQOs) was achieved. If adequate sensitivity was achieved, and bias is "under control," the key issues revolve around whether an adequate number of samples was obtained, given the observed measurement, spatial and temporal variability, and given the actual magnitude of the measurements made (relative to levels of concern). If a data collection effort fails to generate adequate data, then interest in DQIs is heightened, especially if there is a desire to "diagnose" which assumptions proved valid or invalid. Having documented the basis for the design based on DQIs, a quick analysis of where the system failed to perform adequately can be conducted, and a determination of what data will be required to support the decision can be made.

CHAPTER 7

REFERENCES

ANSI/ASQC E4-1994, *Specifications and Guidelines for Environmental Data Collection and Environmental Technology Program*. 1994. American Society for Quality. Milwaukee, WI.

ASTM (American Society for Testing and Materials). 2001a. *D6091-97 Standard Practice for 99%/95% Interlaboratory Detection Estimate (IDE) for Analytical Methods with Negligible Calibration Error*. West Conshohocken, PA.

ASTM (American Society for Testing and Materials). 2001b. *D6512-00 Standard Practice for Interlaboratory Quantitation Estimate*. West Conshohocken, PA.

40 CFR 136, *Code of Federal Regulations, Appendix B, Definition and Procedure for the Determination of the Method Detection Limit - Revision 1.11*. Washington, DC.

50 FR 46906, 1985, *50, Federal Register, Part 46906.* November 13, 1985.

Box, G. E. P., W. G. Hunter, and J. S. Hunter. 1978. *Statistics for Experimenters*. John Wiley. New York.

Carden, K.M., 1998, Method Detection Limit Survey Results and Analysis. Wisconsin Dept of Natural Resources, Laboratory Certification Program. PUBL-SS-930-98.

Chai, E. 1996. *A Systematic Approach to Representative Sampling in the Environment*, in "Sampling Environmental Media, ASTM STP 1282". James Howard Morgan Editor, American Society for Testing and Materials. Philadelphia, PA.

Clayton, C. A., J. W. Hines, and P. D. Elkins. 1987. *Detection Limits with Specified Assurance Probabilities*. Analytical Chemistry, Vol 59, pp. 2506-2514.

Cleveland, W. S. 1993. *Visualizing Data.* Hobart Press. Summit, New Jersey.

Cochran, W. G. and D. R. Cox. 1957. *Experimental Designs*, second edition. John Wiley. New York.

Currie, L. A. 1995. *Nomenclature in Evaluation of Analytical Methods Including Detection and Quantification Capabilities.* Pure and Applied Chemistry, Vol 67, No. 10, pp. 1699-1723.

Gibbons, R. D. 1994. *Statistical Methods for Groundwater Monitoring*. Wiley, NY.

Gibbons, R. D., D. E. Coleman, and R. F. Maddalone. 1997. *An Alternative Minimum Level Definition for Analytical Quantification.* Environmental Science and Technology, Vol 31, No. 7, pp. 2071-2077.

Gibbons, R. D., D. E. Coleman, and R. F. Maddalone. 1997. *Response to Comment on "An Alternative Minimum Level Definition for Analytical Quantification."* Environmental Science and Technology, Vol 31, No.12, pp. 3729-3731.

Gibbons, R. D., D. E. Coleman, and R. F. Maddalone. 1998. *Response to Comment on "An Alternative Minimum Level Definition for Analytical Quantification."* Environmental Science and Technology, Vol 32, No.15, pp. 2349-2353.

Gilbert, R. O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinholt Company. New York.

Glaser, J. A., D. L. Foerst, G. D. McKee, S. A. Quave, and W. L. Budde. December 1981. *Trace Analysis for Wastewaters*. Environmental Science and Technology, Vol 15, No. 12, pp. 1426-1435.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. John Wiley and Sons, New York, p. 447.

Hubaux, A. G. and G. Vos. 1970. *Decision and Detection Limits for Linear Calibration Curves*. Analytical Chemistry, Vol 42, No. 8, pp. 849-855.

Kahn, H. D., W. A. Telliard, and C. E. White. 1998. *Comment on "An Alternative Minimum Level Definition for Analytical Quantification."* Environmental Science and Technology, Vol 32, No. 15, pp. 2346-2348.

Kahn, H. D., W. A. Telliard, and C. E. White. 1999. *Response to Comment on "An Alternative Minimum Level Definition for Analytical Quantification."* Environmental Science and Technology, Vol 33, No. 8, p. 1315.

Keith, L. H. 1991. *Environmental Sampling and Analysis: A Practical Guide*. Lewis Publishers.

Kimbrough, D.E and J. Wakakuwa. 1993. *Method Detection Limits in Solid Waste Analysis*. Environmental Science and Technology, Vol 27, No. 13, pp. 2692-2699.

Kimbrough, D. E. and J. Wakakuwa. 1994. *Quality Control Level: An Alternative to Detection Levels*. Environmental Science and Technology, Vol 28, No. 2, pp. 338-345.

Lame, F. P. and P. R. Defize. 1993. *Sampling of Contaminated Soil: Sampling Error in Relation to Sample Size and Segregation*. Environmental Science and Technology, Volume 27, No. 10. The American Chemical Society. Columbus, OH.

Lancaster, V. A. and S. Keller-McNulty. 1998. "A Review of Composite *Refereed* Sampling Methods," *Journal of the American Statistical Association*. 93(443):1216-1230.

Mandel, J. and T. Lashof. 1987. Concepts of Repeatability and Reproducibility. Journal of Quality Technology. Vol 19, No. 1. pp 29-36.

Oblinger Childress, C. J., W. T. Foreman, B. F. Connor, and T. J. Maloney. 1999. New Reporting Procedures Based on Long-Term Method Detection Levels and Some Considerations for Interpretations of Water-Quality Data Provided by the U. S. Geological Survey National Water Quality Laboratory. USGS Open-File Report 99–193. Reston, VA.

Osborne, K. and D. Rocke. 2000. *The Calculation of Detection Limits using a Two Component Error Model and Laboratory QC Data*. Water Environment Laboratory Solutions. Vol 7 No. 4.

Pitard, F. F. 1993. *Pierre Gy's Sampling Theory and Sampling Practice*, second edition. CRC Press. Boca Raton, Florida.

Rigo, G. 1999. *Comment on "An Alternative Minimum Level Definition for Analytical Quantification."* Environmental Science and Technology, Vol 33, No. 8, pp. 1313-1314.

Rosecrance, A. November/December 2000. *The Three "Rs" for Relevant Detection, Reliable Quantitation, and Respectable Reporting Limits.* Environmental Testing and Analysis, Vol 9, No. 5.

Taylor, J. K. 1987. *Quality Assurance of Chemical Measurements*, Lewis Publishers. Chelsea, Michigan.

Skoog, D. A. 1985. *Principles of Instrumental Analysis*, third edition. Saunders College Publishing. Philadelphia, PA.

U.S. Environmental Protection Agency. 1990. *A Rationale for the Assessment of Errors in the Sampling of Soils*. EPA/600/4-90/007. Washington, DC.

U.S. Environmental Protection Agency. 1992. *Methods for the Detection of Organic Compounds in Drinking Water, Supplement II*. EPA/600/R-92/129. Washington, DC.

U.S. Environmental Protection Agency. 1997. *Multi-Agency Radiation Survey and Site Investigation Manual (MARSSIM)*. NUREG-1575, EPA/402/R-97/016. Washington, DC.

U.S. Environmental Protection Agency. 1998a. *Guidance for Quality Assurance Project Plans (EPA QA/G-5)*. EPA/600/R-98/018. Washington, DC.

U.S. Environmental Protection Agency. 2000a. EPA Order 5360.1 A2, *Policy and Program Requirements for the Mandatory Agency-wide Quality System*. Washington, DC.

U.S. Environmental Protection Agency. 2000b. EPA Order 5360 A1, *EPA Quality Manual for Environmental Programs.* Washington, DC.

U.S. Environmental Protection Agency. 2000c. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis (EPA QA/G-9).* EPA/600/R-96/084. Washington, DC.

U.S. Environmental Protection Agency. 2000d. *Guidance for the Data Quality Objectives Process (EPA QA/G-4).* EPA/600/R-96/055. Washington, DC.

U.S. Environmental Protection Agency. 2000e. *Guidance for Choosing a Sampling Design for Environmental Data Collection (EPA QA/G-5S).* Peer Review Draft. Washington, DC.

U.S. Environmental Protection Agency. 2001a. *Guidance on Environmental Data Verificatin and Validation (EPA QA/G-8).* Peer Review Draft. Washington, DC.

U.S. Environmental Protection Agency. 2001b. *Requirements for QA Project Plans (EPA QA/R-5).* EPA/240/B-01/002. Washington, DC.

Wisconsin Department of Natural Resources, 1996. Analytical Detection Limit Guidance and Laboratory Guide for Determining Method Detection Limits. PUBL-TS-056-96.

Zorn, M. E., R. D. Gibbons, and W. C. Sonzogni. 1991. *Evaluation of Approximate Methods for Calculating the Limit of Detection and Limit of Quantification.* Environmental Science and Technology, Vol 33, No. 13, pp. 2291-2295